American Psychological Association

Journal of Educational Psychology

0022-0663/19/$12.00

2019, Vol. 111, No. 2, 189–209
http://dx.doi.org/10.1037/edu0000282

Retrieval Practice & Bloom's Taxonomy: Do Students Need Fact Knowledge Before Higher Order Learning?

Pooja K. Agarwal
Washington University in St. Louis

The development of students' higher order learning is a critical component of education. For decades, educators and scientists have engaged in an ongoing debate about whether higher order learning can only be enhanced by building a base of factual knowledge (analogous to Bloom's taxonomy) or whether higher order learning can be enhanced directly by engaging in complex questioning and materials. The relationship between fact learning and higher order learning is often speculated, but empirically unknown. In this study, middle school students and college students engaged in retrieval practice with fact questions, higher order questions, or a mix of question types to examine the optimal type of retrieval practice for enhancing higher order learning. In laboratory and K-12 settings, retrieval practice consistently increased delayed test performance, compared with rereading or no quizzes. Critically, higher order and mixed quizzes improved higher order test performance, but fact quizzes did not. Contrary to popular intuition about higher order learning and Bloom's taxonomy, building a foundation of knowledge via fact-based retrieval practice may be less potent than engaging in higher order retrieval practice, a key finding for future research and classroom application.

Educational Impact and Implications Statement
This study demonstrates that students' higher order learning increases most from higher order retrieval practice, or no-stakes quizzes with complex materials that engage students in bringing what they know to mind. Although fact quizzes were beneficial for fact learning, they did not facilitate higher order learning, contrary to popular intuition based on Bloom's taxonomy.

Keywords: Bloom's taxonomy, higher order learning, retrieval practice, testing effect, transfer

Supplemental materials: http://dx.doi.org/10.1037/edu0000282.supp

The development of students' higher order learning is a critical component of education. From both an educational perspective and a scientific perspective, it is of practical interest to develop robust strategies that increase higher order learning. For decades, educators and scientists have engaged in an ongoing debate about instructional approaches: Should we build students' foundation of factual knowledge *before* engaging them in higher order learning, or can higher order learning be enhanced *directly* by engaging students in complex instructional techniques during the initial learning process?

Regarding the first approach, many argue that to foster higher order learning, we must focus on and reinforce students' basic knowledge. For instance, cognitive scientist Daniel Willingham wrote, "Factual knowledge must precede skill" (2009, p. 19). Diane Ravitch (2009), an education professor and historian, argued,

> We have neglected to teach [teachers] that one cannot think critically without quite a lot of knowledge to think about. Thinking critically involves comparing and contrasting and synthesizing what one has learned. And a great deal of knowledge is necessary before one can begin to reflect on its meaning and look for alternative explanations.

This view is not new; there has been a long-standing call to spend a substantial amount of instructional time on foundational fact learning spanning more than 100 years (Bartlett, 1958; Bruner, 1977; Hirsch, 1996; James, 1900; Münsterberg, 1909).

On the other hand, many educators hold that less instructional time should be spent on fact learning and more time should be

This article was published Online First June 7, 2018.

Pooja K. Agarwal, Department of Psychological and Brain Sciences, Washington University in St. Louis.

I thank my Ph.D. advisor and dissertation committee chair, Henry L. Roediger, III, for his dedicated mentorship and guidance. I also thank my dissertation committee members for their significant input: David Balota, Mark McDaniel, Michael Strube, Susan Fitzpatrick, and Keith Sawyer. I am grateful to Columbia Middle School teacher Patrice Bain, principal Roger Chamberlain, and the students for their participation. I also thank the Roediger Memory Lab and the Balota Cognitive Psychology Lab, particularly Jason Finley and Geoffrey Maddox, for valuable discussions and assistance with data analyses. This research was supported by the National Science Foundation Graduate Research Fellowship Program, the Harry S. Truman Scholarship Foundation, and the James S. McDonnell Foundation.

Correspondence concerning this article should be addressed to Pooja K. Agarwal, who is now at the Department of Liberal Arts, Berklee College of Music, Boston, MA 02115. E-mail: pooja@poojaagarwal.com

allotted for classroom activities that promote critical thinking, analysis, and inquiry (Cuban, 1984; Dewey, 1916/1944; Kohn, 1999). For example, education professor Jal Mehta (2018) recently argued,

> What if, in science, we taught students the scientific method . . . by having them write junior versions of scientific papers rather than reading from textbooks? . . . [W]hy in school, do we think it has to be dry basics first, and the interesting stuff only later?

According to the National Research Council (1987, p. 8),

> [T]he term "higher order" skills is probably itself fundamentally misleading, for it suggests that another set of skills, presumably called "lower order," needs to come first. This assumption . . . justifies long years of drill on the "basics" before thinking and problem solving are demanded.

Higher Order Learning & Bloom's Taxonomy

What, exactly, is "higher order" learning? Although there are few agreed-upon definitions, higher order learning is frequently classified using *The Taxonomy of Educational Objectives* by Bloom, Engelhart, Furst, Hill, and Krathwohl (1956). The original "Bloom's taxonomy" included six categories of cognitive processes, ranging from simple to complex: *knowledge*, *comprehension*, *application*, *analysis*, *synthesis*, and *evaluation*. Bloom et al. explained that the taxonomy was designed as a step process: to achieve a higher objective or category, one must first master cognitive processes at a lower category. In other words, before comprehension, application, or analysis can take place, a student must first acquire knowledge.

In 2001, Anderson and coauthors of the original taxonomy proposed a revised taxonomy (see Figure 1). The revised taxonomy highlights learning in verb tense: *remember*, *understand* (previously called comprehension), *apply*, *analyze*, *evaluate*, and *create* (previously called synthesis and reordered with evaluation). Within the revised taxonomy, higher order learning is considered to comprise the *apply*, *analyze*, *evaluate*, and *create* categories. On the other hand, "lower order" learning requiring recognition, memory, and comprehension fall under the *remember* and *understand* categories.

Bloom's taxonomy has had a large impact on teacher preparation programs, classroom pedagogy, small- and large-scale assessment programs, and educational research. In part because of its simplicity, Bloom's taxonomy has contributed to the collective notion that foundational knowledge (literally the foundation or base of the pyramid) precedes higher order learning (the categories located higher in the pyramid). For example, educator Doug Lemov (2017) explained,

> Generally when teachers talk about "Bloom's taxonomy," they talk with disdain about "lower level" questions. They believe, perhaps because of the pyramid image which puts knowledge at the bottom, that knowledge-based questions, especially via recall and retrieval practice, are the least productive thing they could be doing in class. No one wants to be the rube at the bottom of the pyramid.

Lemov's comment perfectly illustrates the existing struggle to identify whether knowledge at the base of Bloom's taxonomy is a prerequisite for higher order learning, or simply a nuisance in the way of higher order learning.

In addition to Bloom's taxonomy, how else might we define or classify higher order learning? More recently, a taxonomy by Barnett and Ceci (2002) is being used to classify students' transfer of knowledge across a number of content and context domains, including physical location, type of task, and format. For example, students' "near transfer" along the content or knowledge domain could involve learning how to calculate the Pythagorean theorem and then applying this knowledge with a novel set of numbers. In

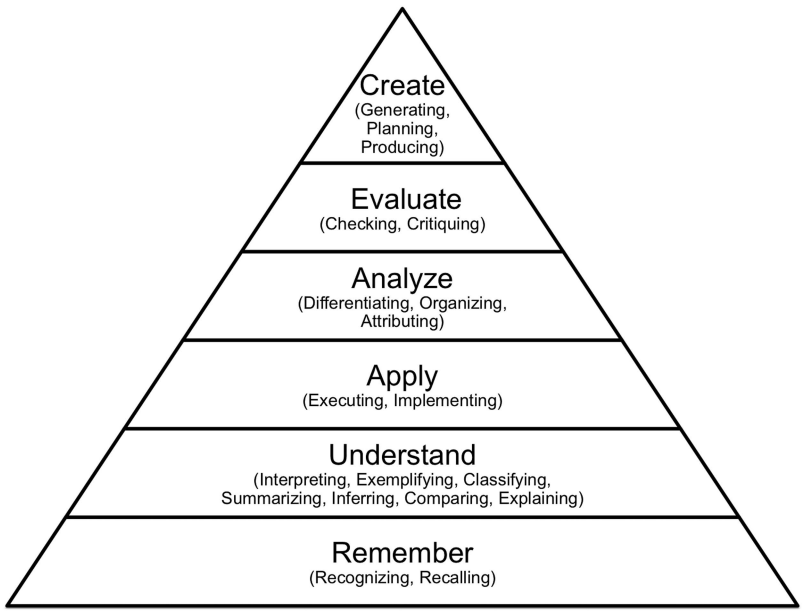

Figure 1. An illustration of the revised Bloom's Taxonomy, based on Anderson et al. (2001).

contrast, "far transfer" along the physical domain could include a medical student learning something from a textbook and then applying it in clinical practice with patients (Pan & Agarwal, 2018; Pan & Rickard, in press).

In terms of higher order learning, we are interested in far transfer, which could occur across content and/or context domains. In addition, Barnett and Ceci (2002) briefly offered a distinction between horizontal and vertical transfer—where horizontal transfer involves two tasks at the same level of complexity and vertical transfer involves learning knowledge that is required for a wide array of tasks that differ in complexity (p. 622, footnote 8). They do not elaborate on the horizontal versus vertical distinction further, unfortunately, but a third domain of "cognitive complexity" may be crucial in identifying the type of higher order learning we seek in educational settings.

Retrieval Practice and Transfer of Knowledge

Based on decades of scientific research, we have identified robust strategies for enhancing students' fact learning and transfer of knowledge, including retrieval practice, spaced practice, and interleaving (Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013; McDaniel, Roediger, & McDermott, 2007; Pashler et al., 2007; Rohrer & Pashler, 2010). One strategy in particular—retrieval practice—dramatically improves long-term learning (Agarwal, Roediger, McDaniel, & McDermott, 2017; Roediger & Karpicke, 2006b). In typical experiments, students study a set of material (e.g., word pairs, foreign language vocabulary words, prose passages), engage in retrieval practice (e.g., via free recall or multiple-choice quizzes), and immediately or after a delay (e.g., ranging from hours, to days, to weeks) they complete a final test. When students engage in retrieval activities that bring knowledge to mind, learning is strengthened by the challenge (for recent meta-analyses, see Adesope, Trevisan, & Sundararajan, 2017; Rowland, 2014).

Specifically, recent research in classroom settings has demonstrated that retrieval practice improves learning for diverse student populations (e.g., middle school students to medical school students), subject areas (e.g., introductory history to CPR skills), and time delays (e.g., from a few days to 9 months; Kromann, Jensen, & Ringsted, 2009; Lyle & Crawford, 2011; Roediger, Agarwal, McDaniel, & McDermott, 2011). In addition, benefits from retrieval practice have been demonstrated in diverse educational settings, including K-12 classrooms (Agarwal, Bain, & Chamberlain, 2012), undergraduate engineering courses (Butler, Marsh, Slavinsky, & Baraniuk, 2014), and medical neurology courses (Larsen, Butler, Lawson, & Roediger, 2013).

Retrieval practice also promotes students' learning of complex materials (Jensen, McDaniel, Woodard, & Kummer, 2014; Karpicke & Aue, 2015; Pyc, Agarwal, & Roediger, 2014; Rawson, 2015; cf., van Gog & Sweller, 2015). In one study examining the impact of retrieval practice on the learning of complex materials, subjects studied bird exemplar picture-family name pairs (e.g., a picture of a bird from the thrasher family) in either a repeated study condition or a repeated quiz condition (Jacoby, Wahlheim, & Coane, 2010). In the repeated study condition, subjects viewed pictures of birds paired with their family names four times. For the quiz condition, subjects studied picture-family name pairs once, followed by three attempts to classify pictures for the eight bird families. On a final test, recognition and classification performance for picture exemplars was greater following repeated quizzes than restudying. The materials in this study were more complex than typical laboratory materials (e.g., word pairs, textbook passages), whereas the cognitive processes comprised the lowest two categories of Bloom's taxonomy: remember and understand.

Moving to complex cognitive processes, recent research provides compelling evidence that retrieval practice is an effective strategy to promote students' transfer of learning (Butler, Black-Maier, Raley, & Marsh, 2017; Carpenter, 2012; Karpicke & Blunt, 2011; Pan & Rickard, in press; Rohrer, Taylor, & Sholar, 2010). Butler (2010) examined students' transfer of knowledge across the content domain described by Barnett and Ceci (2002), using passages drawn from Wikipedia and similar online sources. Subjects read multiple passages and some were followed by repeated studying, whereas others were followed by repeated quizzing. On final tests, repeated quizzes led to greater fact and inferential learning, compared with repeated studying. Importantly, when subjects were given final test questions from a different knowledge domain (e.g., an initial quiz question about the wing structure of bats, in contrast with a final test question about the wing structure of military aircraft), repeated quizzes led to greater transfer performance than repeated studying. In other words, retrieval practice improved students' transfer of content knowledge from bats to airplanes and it also improved students' inferential processing, moving closer toward what we think of as higher order learning in terms of both content and complexity.

In a real-world demonstration of the benefits of retrieval practice on transfer, McDaniel, Thomas, Agarwal, McDermott, and Roediger (2013, Experiment 2) conducted an experiment in an 8th grade science classroom in which retrieval practice included fact and application questions (compared with no quizzing) and the final tests also included fact and application questions. Of interest was which type of retrieval practice would enhance final test performance two weeks later, particularly on final application questions. Retrieval practice significantly increased final test performance regardless of format, and the application quiz condition was most beneficial for application test performance. In other words, when it comes to promoting complex learning, such as students' application of knowledge, engaging in complex retrieval practice was more beneficial than starting with basic facts and definitions. This finding supports the notion that higher order learning can be enhanced directly by higher order retrieval practice, one of the first such demonstrations in an authentic classroom setting. The study by McDaniel et al. (2013) did not examine transfer across content or context domains; however, the materials in the study extended students' cognitive complexity from the remember category to the apply category—the next step higher in Bloom's taxonomy (see Figure 1).

Critically, how far "up the pyramid" can we move student learning, based on Bloom's taxonomy? Should we first build students' foundation of knowledge or can we skip straight ahead to complex retrieval practice during initial learning? The primary purpose of this study was to examine the intuition that "factual knowledge must precede skill" (Willingham, 2009). By using complex passages, quizzes, and tests, I aimed to shed light on the relationship between fact learning and higher order learning—a critical distinction with practical and theoretical implications for research and instruction.

Theoretical Rationale

Three theoretical frameworks were utilized to explore the relationship between fact knowledge and complex higher order learning: the desirable difficulties framework, the transfer appropriate processing framework, and what I refer to as the "foundation of factual knowledge framework." Importantly, these theories are not mutually exclusive; rather, the three frameworks suggest both similar and different results depending on the experimental condition in the present study (see Table 1). By teasing apart these theoretical similarities and differences in the current study, optimal strategies to promote higher order learning can be developed and implemented in authentic educational settings.

Desirable Difficulties Framework

According to the desirable difficulties framework, mental processes that are challenging and effortful typically increase retention and application of knowledge (Bjork, 1994). Exposure to "desirable difficulties," particularly via retrieval activities, produces a struggle to bring information to mind that improves learning and delays forgetting. In addition to retrieval practice, a number of other strategies enhance learning to a greater extent than their less effortful comparisons, such as spaced practice instead of massed practice (Rohrer & Taylor, 2006) and interleaved practice instead of blocked practice (Kornell & Bjork, 2008; Taylor & Rohrer, 2010; see Dunlosky et al., 2013 for a review).

Furthermore, difficult retrieval practice may benefit delayed test performance to a greater extent than easier retrieval practice (Bjork, 1994; Gardiner, Craik, & Bleasdale, 1973; Pyc & Rawson, 2009). For example, Kang, McDermott, and Roediger (2007) found that retrieval practice with short answer quizzes enhanced final test performance to a greater extent than retrieval practice with multiple-choice quizzes. Kang et al. concluded that short answer quizzes may engage deeper recollective processing and greater retrieval effort compared with multiple-choice quizzes, thereby enhancing performance on both final short answer and multiple-choice tests. These findings support the notion that retrieval practice with challenging questions (short answer or application) can benefit performance on less and more demanding criterial tests (both multiple-choice and short answer, or definition and application; cf., McDermott, Agarwal, D'Antonio, Roediger, & McDaniel, 2014).

Regarding the present study, the desirable difficulties framework suggests that delayed test performance will be greater following retrieval practice compared with restudying or no quizzes, consistent with previous research (e.g., Agarwal, Karpicke, Kang, Roediger, & McDermott, 2008; Callender & McDaniel, 2009; Carrier & Pashler, 1992). Initial higher order quizzes may serve as a desirable difficulty, enhancing performance on both fact and higher order tests (see Table 1). Meanwhile, mixed quizzes with both fact and higher order questions (in Experiments 2 and 3) may pose an additional challenge because the difficulty of questions varies within quizzes, leading to even greater benefits on final test performance than higher order quizzes.

Transfer Appropriate Processing

According to the transfer appropriate processing framework, final performance is greatest when encoding processes engaged during learning match retrieval processes engaged during testing (McDaniel, Friedman, & Bourne, 1978; Morris, Bransford, & Franks, 1977). This framework is often cited as another explanation for why retrieval practice enhances long-term retention—by engaging in retrieval practice, students match their initial processing to the type of processing required at test (Roediger & Karpicke, 2006a).

In the classroom study described earlier, McDaniel et al. (2013) demonstrated that challenging application quizzes promoted learning for both basic definition and complex application tests. They

Table 1
Theoretical Predictions and Obtained Results of Initial Retrieval Practice Condition on Final Test Performance, Compared With a No Quiz Condition

Retrieval practice	Final test	Desirable difficulties	Transfer appropriate processing	Foundation of factual knowledge	Obtained results
Fact quiz	Fact test	++	++	++	E1: ++ E2: ++ E3: N/A
Higher order quiz	Higher order test	++	++	0	E1: ++ E2: ++ E3: ++
Fact quiz	Higher order test	+	0	++	E1: 0 E2: 0 E3: N/A
Higher order quiz	Fact test	++	0	0	E1: 0 E2: 0 E3: N/A
Mixed quiz	Fact test	++	+	+	E1: N/A E2: ++ E3: ++
Mixed quiz	Higher order test	++	+	+	E1: N/A E2: ++ E3: ++

Note. Facilitation: ++; partial facilitation: +; no prediction or effect: 0; not applicable: N/A.

also found results consistent with the transfer appropriate processing framework: Across two experiments, test performance was greatest when quiz questions matched the final test format, compared with no quizzes or a quiz-test mismatch. For example, student performance was greater on the application test following application quizzes instead of definition quizzes.

Regarding the present study, the transfer appropriate processing framework suggests that delayed test performance will be optimal when initial quiz and final test formats match (see Table 1). Specifically, retrieval practice with fact questions should benefit delayed fact performance, and retrieval practice with higher order questions should benefit delayed higher order performance (compared with no quizzes or restudying). Further, a match between quiz and test questions may promote performance to a greater extent than a mismatch (fact quiz-higher order test or higher order quiz-fact test).

Foundation of Factual Knowledge Framework

In accordance with the original Bloom's taxonomy, the "foundation of factual knowledge" framework suggests that we must focus on and reinforce basic factual knowledge before we can foster students' higher order learning (Brown, Roediger, & McDaniel, 2014). For instance, Willingham (2009) argued that when students practice facts until they are memorized, students can more easily apply their deeper knowledge to higher order learning. Retrieval practice of facts and procedures (e.g., applying the distributive property in algebra) may make recall automatic, thereby requiring less effort and capacity from working memory. Subsequently, this process may enable students to use additional working memory capacity in more complex learning situations (Agarwal, Finley, Rose, & Roediger, 2017). Consistent with these arguments, cognitive load theory posits that if cognitive demands are reduced or diminished, subsequent learning may increase (Plass, Moreno, & Brünken, 2010; Sweller, 2010; van Gog & Sweller, 2015). As such, facilitation of fact learning via retrieval practice may reduce cognitive demands and enhance final higher order test performance, consistent with the foundation of factual knowledge framework. Mental effort ratings (Paas, 1992) were collected in the present study to elucidate this possibility.

Note that Willingham (2009) also distinguished between rote knowledge and integrated or connected knowledge, the latter of which may be required for higher order learning (see also Ausubel, 1961/1965; Ausubel, Novak, & Hanesian, 1978). For instance, knowing the basic fact that George Washington was the first president of the United States may not lead to a deeper understanding of United States civics or government. Willingham argued that, instead, the learning of connected knowledge (e.g., presidents are leaders who make important decisions) can facilitate deep knowledge (e.g., if George Washington was the first president, he must have made many important decisions) to construct a rich understanding of a topic. In other words, simply learning isolated facts without connecting them to a deeper understanding of a topic may not benefit students' higher order learning.

To directly examine the foundation of factual knowledge framework, experiments included fact-based retrieval practice and delayed higher order tests. All fact questions in the present study were developed to encompass key concepts or ideas from passages (i.e., integrated knowledge, such as the goal of welfare programs), rather than details such as names, dates, vocabulary words, definitions, and so forth (such as the year in which the Temporary Assistance for Needy Families program was initiated). In keeping with Willingham (2009) and the foundation of factual knowledge framework, if an understanding of integrated concepts is required for higher order learning, then delayed higher order test performance should benefit from integrated fact quizzes (see Table 1). Note that this prediction contradicts the transfer appropriate processing framework, which suggests that the match between higher order quizzes and a delayed higher order test will result in greater test performance than the mismatch between fact quizzes and a delayed higher order test.

Introduction to Experiments

The relationship between fact learning and higher order learning is often speculated, but empirically unknown. To maximize learning while also experimentally manipulating cognitive complexity, I used a retrieval practice paradigm with educational materials that engaged students in lower order (e.g., remembering and understanding) and higher order (e.g., analyzing, evaluating, and creating) cognitive processes (see Figure 1). I also conducted a conceptual replication in an authentic K-12 classroom in which retrieval practice was embedded in students' and teachers' daily activities, fluctuating schedules, and standard lessons (Experiment 3). Across three experiments, if initial fact learning increases delayed higher order learning, findings will shed light on a longstanding theoretical and practical debate—how best to achieve the "holy grail" of the highest levels in Bloom's taxonomy.

Experiment 1

In Experiment 1, college students participated in four retrieval practice conditions: a study once condition, a study twice condition, study once followed by a fact quiz, and study once followed by a higher order quiz. After two days, students completed fact and higher order tests for each condition.

Method

Participants. Forty-eight college students (M age = 20.58 years, 29 females) were recruited from the Department of Psychology human subjects pool. Subjects received either credit toward completion of a research participation requirement or cash payment ($25). Analyses were conducted only after data from 48 subjects were collected, a sample size determined at the outset of the study using a power analysis with an assumed effect size of $d = 0.5$.

Design. A 4 × 2 within-subject design was used, such that four retrieval practice conditions (study once, study twice, fact quiz, higher order quiz) were crossed with two delayed test types (fact test, higher order test). Eight passages, two per retrieval practice condition, were presented in the same order for all subjects, but the order in which the conditions occurred was blocked by retrieval practice condition and counterbalanced using a Latin Square (see Appendix A). Retrieval practice conditions appeared once in every ordinal position and were crossed with the two types of final tests, creating eight counterbalancing orders. Six subjects were randomly assigned to each of the eight orders. After a 2-day

delay (i.e., 48 hr later), subjects completed one test type (a fact test or a higher order test) per passage, with tests presented in the same order in which passages were encountered during Session 1.

Materials. Eight passages were adapted from eight books included in the "Taking Sides" McGraw-Hill Contemporary Learning Series (http://www.mhcls.com; Daniel, 2006; Easton, 2006; Madaras & SoRelle, 1993; Moseley, 2007; Noll, 2001; Paul, 2002; Rourke, 1987). Each passage was approximately 1,000 words in length (M = 1,006 words, range = 990 to 1016 words), with half of each passage presenting one viewpoint of a controversial topic, and the remaining half of each passage presenting the opposite viewpoint (all passages are included in the online supplementary material). For example, a passage entitled, "Does welfare do more harm than good?" was adapted from *Taking Sides: Clashing Views on Controversial Social Issues* (Finsterbusch & McKenna, 1984), for which 500 words were drawn from the book to describe a "yes" argument and approximately 500 words were used to describe a "no" argument.

For Session 1, eight four-alternative multiple-choice fact questions and eight four-alternative multiple-choice higher order questions were developed for each passage (see Appendix B). For each question type, approximately four questions pertained to the "yes" argument and approximately four questions pertained to the "no" argument. For Session 2, all question stems were rephrased and multiple-choice alternatives were randomly reordered, but alternatives were not rephrased. Across Sessions 1 and 2, regardless of counterbalancing order, the correct multiple-choice alternative appeared in every position (1, 2, 3, or 4) an equal number of times.

For fact questions, broad ideas stated in the passages were tested to measure subjects' overall understanding of the content. For example, a fact question from the "Does welfare do more harm than good?" passage included:

> Which is the primary reason the "yes" author is against welfare programs?
>
> 1. Welfare programs do not benefit recipients or taxpayers
> 2. Welfare programs create dependence for recipients
> 3. Welfare programs are too expensive for taxpayers
> 4. Welfare programs are not the government's responsibility

The correct answer for this fact question is alternative #1, and the answer was stated directly in the passage. Critically, all fact questions in the present study were designed to encompass key concepts or ideas from passages, rather than details such as names, dates, vocabulary words, definitions, and so forth (e.g., the definition of welfare).

Higher order questions were developed in accordance with the *apply*, *analyze*, *evaluate*, and *create* categories of Anderson et al.'s (2001) revised Bloom's taxonomy (see Figure 1). For *apply* questions, subjects were asked about a new situation or problem that was related to a broad idea that was stated in the passage. For example, an *apply* question from the welfare passage included:

> What type of society would the "yes" author expect if there were no welfare programs in the future?
>
> 1. A society in which all individuals are self-reliant and independent
> 2. A society in which there would be no role for the government
> 3. A society in which no one would be required to pay taxes
> 4. A society in which all individuals are treated equally

The correct answer for this *apply* question is alternative #1 and it could be inferred from the passage, but it was not stated explicitly.

For *analyze* questions, subjects were asked to differentiate the authors' arguments; they were presented with a statement and asked which author (the "yes" author, the "no" author, both authors, or neither author) would agree or disagree with the statement. For example, an *analyze* question included:

> Which author would agree with the following statement? "It is honorable for the government to help society."
>
> 1. The "yes" author
> 2. The "no" author
> 3. Both authors
> 4. Neither author

The correct answer for this *analyze* question is alternative #3.

For *evaluate* questions, subjects were asked to check or critique an author's argument by selecting a statement (which was not presented in the passage) that most accurately summarized the author's argument. For example, an *evaluate* question included:

> Which statement is an accurate evaluation or summary of the "yes" author's views?
>
> 1. Welfare programs can never work, because they are always too expensive
> 2. Welfare programs are harmful, because they make bad situations even worse
> 3. Welfare programs waste taxpayer money on people who do not really need help
> 4. Welfare programs could work, but they rarely meet the needs of the people

The correct answer for this *evaluate* question is alternative #4.

Lastly, for *create* questions, subjects were asked to plan or predict an outcome for a novel situation that was not stated in the passage; thus, the author's potential reaction must be generated based on information presented in the passage. For example, a *create* question included:

> How do you predict the "yes" author would react if he or she became unemployed and needed welfare assistance?
>
> 1. The "yes" author might accept government assistance, but would seek help from local organizations first
> 2. The "yes" author would not accept government assistance, but would try to find a new job

3. The "yes" author might accept government assistance, but would try to find a new job first

4. The "yes" author would not accept government assistance, but would seek help from local organizations

The correct answer for this *create* question is alternative #2. All reading passages and questions developed are included in the online supplementary materials (see Appendix B for sample questions).

Procedure. Subjects were tested in small groups (up to five people) using E-Prime 2.0 software (Schneider, Eschman, & Zuccolotto, 2007). At the beginning of Session 1, subjects were instructed that they would be reading passages and taking multiple-choice tests. Subjects were presented with a sample passage about the Nicaraguan Contras, drawn from the same book series as experimental materials, for 20 seconds. They were instructed to familiarize themselves with the computer program's scrolling feature (viewing the entire passage using the up and down keys on the keyboard) without worrying about reading the passage. Next, subjects were presented with a practice test of two 4-alternative multiple-choice questions (self-paced) that asked subjects whether they turned off their cell phone and whether they could return in two days; in other words, subjects did not receive sample test questions related to the sample passage. After responding to the two practice questions, subjects were asked to make a mental effort rating on a nine-point scale (adapted from Paas, 1992), which was followed by 10 seconds of feedback, to acclimate subjects to the experimental procedure.

After the instruction phase during Session 1, subjects completed two blocks: First, subjects read all eight passages, presented in the same order for all subjects. Second, subjects completed four within-subject conditions (two per passage), blocked by retrieval practice condition (see Appendix A). In other words during the second block, subjects did not reencounter two of the passages (in the study once condition), they read two of the passages for a second time (in the study twice condition), they completed quizzes with fact questions for two of the passages, and they completed quizzes with higher order questions for the remaining two passages.

During 6-min study periods, each passage was presented in its entirety on the computer screen and subjects were able to scroll up and down to read the complete text at their own pace. Subjects were asked to study the passage during the time allotted and after six minutes, the computer moved on to the next passage (during the first reading block) or to the appropriate condition (during the second condition block following the restudy condition). Subjects' keyboard presses were recorded during study periods to ensure that all subjects scrolled appropriately through the passages from the top of the passage to the bottom.

During self-paced initial quiz periods, multiple-choice questions (blocked by passage) were presented one at a time, in a different random order for each subject. Subjects were asked to type a number (1, 2, 3, or 4) corresponding to the multiple-choice alternative (forced choice). As soon as subjects responded to each question, the computer moved on (i.e., subjects were not allowed to change their answer) and subjects were asked to estimate the mental effort required ("How much mental effort did you invest in this question") on a 9-point scale (adapted from Paas, 1992) by typing a number corresponding to the rating. After subjects rated their mental effort, the computer presented immediate feedback for 10 seconds by displaying the word "CORRECT" or "INCORRECT" corresponding to subjects' response, while also displaying the original question and the correct answer (without incorrect multiple-choice lures). After 10 seconds, the computer moved on to the next question. In other words, multiple-choice question responses were self-paced, whereas feedback was experimenter-controlled and presented for 10 seconds per item. In summary, subjects completed a question, provided a mental effort rating, and then viewed feedback, followed by the next question. After each passage and quiz (regardless of condition), subjects received a 15 second break, during which the computer screen displayed, "Please clear your mind and wait for the computer to move on." Then, the computer moved on to the next passage or condition, according to subjects' counterbalancing order.

After two days, subjects returned for Session 2 and completed multiple-choice fact tests for four of the passages and multiple-choice higher order tests for the other four passages. Testing procedures outlined above for Session 1 were followed during Session 2, except subjects did not receive feedback during Session 2.

In sum, subjects participated in four within-subject retrieval practice conditions, crossed with two delayed test types. Dependent variables measured included accuracy on test questions, response times for test questions, mental effort ratings for test questions, and response times for mental effort ratings. The entire procedure lasted approximately two and a half hours across the two sessions. At the end of the experiment, subjects were debriefed and thanked for their time.

Results

All results in the current study were significant at an alpha level of .05. A Greenhouse-Geisser correction was applied to analyses of variance (ANOVAs) when the sphericity assumption was violated (Greenhouse & Geisser, 1959) and a Bonferroni correction for multiple comparisons was applied to p values from t tests by multiplying the p value by the number of comparisons (Rice, 1989). Effect sizes reported include partial eta-squared (η_p^2) for ANOVAs (Pearson, 1911; Pierce, Block, & Aguinis, 2004) and Cohen's d for t tests (Cohen, 1988). Error bars in figures represent 95% confidence intervals, specifically calculated for within-subject designs according to methods described by Cousineau (2005) and Morey (2008). Data from response times and mental effort ratings did not contribute to the overall findings from the present study, as discussed in the General Discussion. Thus, these data are not reported and are available upon request.

Initial quiz performance. Initial performance on the fact quiz and the higher order quiz is displayed in Table 2. As predicted, initial performance was greater on the fact quiz (59%) compared with performance on the higher order quiz (47%), likely because of item difficulty, confirmed by a one-way ANOVA on initial performance, $F(1, 47) = 12.62$, $p = .001$, $\eta_p^2 = .21$.

Final test performance. Final test performance on rephrased questions for the four retrieval practice conditions is displayed in Table 2 and Figure 2. Reliability (Cronbach's alpha) was .501 for final fact test performance and .467 for final higher order performance. When collapsed over final test type, delayed test performance was lower for the study once (49%) and study twice (51%)

Table 2
Initial Quiz and Delayed Test Performance (Proportion Correct) as a Function of Retrieval Practice Condition From Experiment 1

Condition	Initial quiz	Final fact test	Final higher order test	Delayed average
Study once		.54 (.21)	.44 (.18)	.49
Study twice		.54 (.21)	.49 (.19)	.51
Fact quiz	.59 (.17)	.78 (.19)	.46 (.22)	.62
Higher order quiz	.47 (.15)	.53 (.21)	.72 (.21)	.62
Average	.53	.60	.53	

Note. Standard deviations are displayed in parentheses.

conditions, and greater for the fact quiz (62%) and higher order quiz (62%) conditions. When collapsed over initial learning condition, overall performance on the final fact test was greater (60%) than performance on the final higher order test (53%), likely due to item difficulty. A 4 (retrieval practice condition: study once, study twice, fact quiz, higher order quiz) × 2 (delayed test type: fact, higher order) repeated measures ANOVA on delayed test performance indicated a main effect of retrieval practice condition, $F(3, 141) = 13.47$, $p < .001$, $\eta_p^2 = .22$, a main effect of delayed test type, $F(1, 47) = 12.47$, $p = .001$, $\eta_p^2 = .21$, and a significant interaction between retrieval practice condition and test type, $F(3, 141) = 27.39$, $p < .001$, $\eta_p^2 = .37$.

Regarding delayed performance on the fact test, post hoc *t* tests confirmed a significant effect of retrieval practice, such that final fact test performance for the fact quiz condition (78%) was significantly greater than final fact test performance for the study once (54%) and study twice (54%) conditions, $t(47) = 5.96$, $p < .001$, $d = 1.23$ and $t(47) = 6.63$, $p < .001$, $d = 1.24$, respectively. However, final fact test performance was similar for the higher order quiz condition (53%) compared with the study once and study twice conditions, ts < 1, indicating that retrieval practice with higher order questions did not benefit final fact test performance. In other words, an initial fact quiz improved final fact test performance (78%) to a much greater degree than an initial higher order quiz (53%), $t(47) = 6.93$, $p < .001$, $d = 1.29$. There was no effect of restudying when comparing final fact test performance for the study once and study twice conditions, $t < 1$.

Regarding delayed performance on the higher order test, post hoc *t* tests also confirmed a significant effect of retrieval practice, such that final higher order test performance for the higher order quiz condition (72%) was significantly greater than final higher order test performance for the study once (44%) and study twice (49%) conditions, $t(47) = 8.17$, $p < .001$, $d = 1.39$ and $t(47) = 5.31$, $p < .001$, $d = 1.12$, respectively. However, final higher order test performance was similar for the fact quiz condition (46%) compared with the study once and study twice conditions, ts < 1, indicating that retrieval practice with fact questions did not benefit final higher order test performance. In other words, an initial higher order quiz improved final higher order test performance (72%) to a much greater degree than an initial fact quiz (46%), $t(47) = 6.73$, $p < .001$, $d = 1.21$. Again, there was no effect of restudying on final higher order test performance when comparing the study once and study twice conditions, $t(47) = 1.40$, $p > .05$.

In sum, initial retrieval practice enhanced final test performance, but only when the initial quiz type (fact or higher order) matched the final test type (fact or higher order, respectively). For these congruent conditions, performance was marginally greater for the fact quiz-fact test condition (78%) than for the higher order quiz-higher order test condition (72%), $t(47) = 1.94$, $p = .059$, $d = 0.32$, though this difference may be due to relative difficulty between fact versus higher order test questions.

Discussion

In Experiment 1, retrieval practice with higher order questions greatly improved delayed higher order test performance by 23–28% (compared with studying once or twice, Figure 2). Consistent with prior research, retrieval practice with fact questions also improved delayed fact performance by 24%. When the type of initial quizzes matched the type of final test, even when final test questions were rephrased, retrieval practice yielded comparable benefits on performance for both fact and higher order learning. Thus, results from Experiment 1 are consistent with predictions based on the transfer appropriate processing framework (see Table 1).

Critically, retrieval practice with fact questions did not enhance delayed higher order test performance, contrary to the foundation of factual knowledge framework. In addition, retrieval practice with higher order questions did not enhance delayed fact test performance, contrary to the desirable difficulty framework. There were no benefits from restudying on delayed test performance, even when the first and second study periods were spaced over time, replicating previous findings (Agarwal et al., 2008; Callender & McDaniel, 2009; Carrier & Pashler, 1992; Karpicke & Roediger, 2007; Roediger & Karpicke, 2006b; Wheeler, Ewers, & Buonanno, 2003).

In sum, there were no benefits of initial fact quizzes on delayed higher order test performance, no benefits of initial higher order quizzes on delayed fact test performance, and no benefit of restudying on delayed test performance, regardless of test type. Of

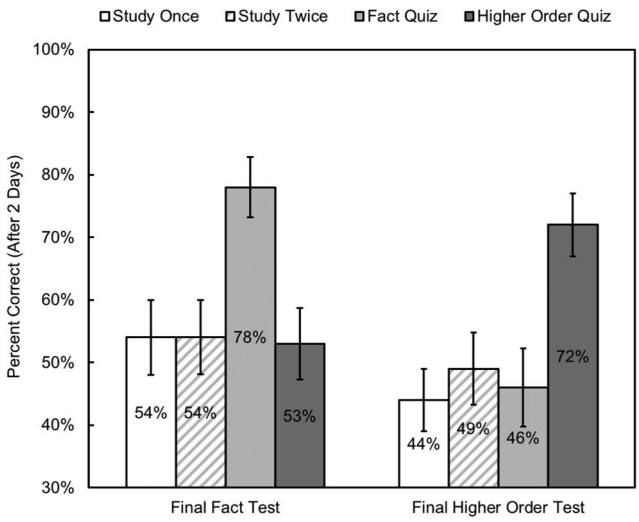

Figure 2. Delayed test performance (proportion correct after two days) as a function of retrieval practice condition from Experiment 1. Errors bars represent 95% confidence intervals.

the three theoretical frameworks, results from Experiment 1 most closely align with the transfer appropriate processing framework.

Experiment 2

Experiment 2 was designed to examine whether a mix of both fact and higher order question types used during retrieval practice would be more beneficial for enhancing delayed test performance than using one type of question during initial retrieval practice. Considering the results from Experiment 1 where fact quizzes did not enhance delayed higher order test performance and higher order quizzes did not enhance delayed fact test performance, retrieval practice with both types of questions during initial learning may benefit delayed test performance, regardless of test type.

When receiving a fact quiz and a higher order quiz in close succession during the first session (i.e., in a mixed condition), students may be more likely to activate elaborative information (Carpenter, 2009), create more robust mediators or connections between concepts (Pyc & Rawson, 2010), and subsequently transfer knowledge from initial learning to a delayed higher order test. In addition, providing students with prompts or hints to transfer their knowledge to novel situations enhances their subsequent transfer performance (Gick & Holyoak, 1980), thus a mixed quiz condition may prompt students to transfer their knowledge during initial learning and also on delayed tests (see also Butler et al., 2017; Pan & Rickard, in press). From an instructional standpoint, recent depictions of Bloom's taxonomy in the shape of a wheel or web eliminate the typical hierarchical structure in favor of an integrated approach (see Knapp, 2016, for examples). Thus, mixed complexity during initial learning may serve as an effective alternative to the prevalent "factual knowledge first" standpoint.

In Experiment 2, subjects participated in four retrieval practice conditions, each after studying a passage once: they completed one higher order quiz, two higher order quizzes, two fact quizzes, and two mixed quizzes. After two days, subjects completed fact and higher order tests for each condition. In Experiment 1, no-quiz and restudy conditions did not improve final test performance. Thus, these conditions were not included in Experiment 2. Instead, the comparisons of interest in Experiment 2 were the optimal combination of quizzes for improving delayed fact and higher order learning—namely, two fact quizzes, two higher order quizzes, or a mix of quizzes.

Given the results from Experiment 1, it was expected that the higher order quiz (once or 1X) and higher order quiz (twice or 2X) conditions would benefit delayed higher order test performance, and that the fact quiz (twice or 2X) condition would benefit delayed fact test performance. Regarding one quiz versus two quizzes, it was predicted that two higher order quizzes would provide an additional benefit to delayed higher order test performance compared with one higher order quiz. On the other hand, this additional benefit may be due to reexposure to the same item twice; that is, question stems were only rephrased between the initial and final sessions, not between the first and second initial quizzes.

Regarding the mixed quiz condition (2X, with one fact and one higher order quiz), it may be the case that including both question types provides students with "the best of both worlds"—the mixed quiz condition could enhance delayed performance on both types of tests compared with the nonmixed retrieval practice conditions (see Table 1). In line with the transfer appropriate processing framework, engaging in both types of processing during initial learning may have the greatest overlap in processing required for the two final test types, enhancing delayed performance. At the same time, one quiz of each format (in the mixed quiz condition) may not prove as potent as two quizzes of the same format; therefore, it was unclear whether the mixed quiz (2X) condition would provide a smaller or larger benefit to delayed test performance compared with the fact (2X) and higher order (2X) quiz conditions.

Method

Participants. Forty-eight college students (M age = 20.04 years, 31 females) were recruited from the Department of Psychology human subjects pool. Subjects received either credit toward completion of a research participation requirement or cash payment ($25). Subjects who participated in Experiment 2 did not participate in Experiment 1. Analyses were conducted only after data from 48 subjects were collected, a sample size determined at the outset of the study using a power analysis with an assumed effect size of $d = 0.5$.

Design. A 4 × 2 within-subject design was used, such that four retrieval practice conditions [higher order quiz (1X), higher order quizzes (2X), fact quizzes (2X), mixed quizzes (2X)] were crossed with two delayed test types (fact test, higher order test). Eight passages, two per retrieval practice condition, were presented in the same order for all subjects. The order in which the conditions occurred was blocked by retrieval practice condition and counterbalanced using a Latin Square (see Appendix A). Retrieval practice conditions appeared once in every ordinal position and were crossed with the two types of final tests, creating eight counterbalancing orders. Six subjects were randomly assigned to each of the eight orders. Specifically for the mixed quiz (2X) condition, subjects completed a fact quiz followed by a higher order quiz, or they completed a higher order quiz followed by a fact quiz. Order of quizzes in the mixed quiz condition was counterbalanced equally across subjects (see Appendix A).

After a 2-day delay (i.e., 48 hr later), subjects completed one test type (a fact test or a higher order test) per passage. Tests were presented in the same order in which passages were encountered during Session 1.

Materials. The same materials from Experiment 1 were used in Experiment 2 (see Appendix B for sample questions).

Procedure. The same procedures used in Experiment 1 were used in Experiment 2, except that subjects completed three blocks during Session 1: First, subjects read all eight passages, presented in the same order for all subjects. Second, subjects completed the first quiz block with eight quizzes (one quiz per passage, presented in the same order as passages during the reading block). Third, subjects completed a second quiz block with six quizzes [one quiz per passage, except for passages in the higher order quiz (1X) condition, again presented in the same order]. After two days, subjects returned for Session 2 and completed multiple-choice fact tests for four of the passages and multiple-choice higher order tests for the other four passages.

In sum, subjects participated in four within-subject retrieval practice conditions, crossed with two delayed test types. Dependent variables measured included accuracy on test questions, re-

sponse times for test questions, mental effort ratings for test questions, and response times for mental effort ratings. The entire procedure lasted approximately two and a half hours across the two sessions. At the end of the experiment, subjects were debriefed and thanked for their time.

Results

Data from response times and mental effort ratings did not contribute to the overall findings from the present study, as discussed in the General Discussion. Thus, these data are not reported and are available upon request.

Initial quiz performance. Initial performance during the first and second quiz blocks is displayed in Table 3. There was no effect of counterbalancing order on initial quiz performance for the mixed quiz (2X) condition (fact-higher order or higher order-fact), $F < 1$, therefore means were collapsed over initial order for subsequent analyses (see Table 3 for the complete set of means). For the first quiz block, initial performance was greatest for the fact quiz (2X) condition (57%), followed by initial performance for the mixed quiz (2X, 52%, collapsed over quiz order), higher order quiz (2X, 49%), and higher order quiz (1X, 47%) conditions, respectively. For the second quiz block, initial performance was again greatest for the fact quiz (2X) condition (91%), followed by the higher order quiz (2X, 83%) and mixed quiz (2X, 53%, collapsed over quiz order) conditions.

A 3 [retrieval practice condition: higher order quiz (2X), fact quiz (2X), mixed quiz (2X)] × 2 (quiz block: first, second) repeated measures ANOVA on initial performance revealed a significant main effect of retrieval practice condition, $F(2, 94) = 64.27$, $p < .001$, $\eta_p^2 = .58$, a significant main effect of quiz block, $F(1, 47) = 356.69$, $p < .001$, $\eta_p^2 = .88$, and a significant interaction between retrieval practice condition and quiz block, $F(2, 94) = 42.77$, $p < .001$, $\eta_p^2 = .48$. As displayed in Table 3, the higher order quiz (2X) and fact quiz (2X) conditions resulted in a similar increase in performance from the first quiz block to the second quiz block (34% for each condition).

Performance in the mixed quiz (2X) condition, on the other hand, remained relatively constant across quiz blocks (see Table 3). Note that performance for each quiz block includes subjects' performance on both types of quizzes (fact and higher order). This finding suggests a replication of Experiment 1, namely that retrieval practice on one quiz format did not benefit performance on a second quiz of a different format, even in close succession during the first session—performance on the second quiz in the mixed condition was similar to performance on the first quiz of the same type in the fact quiz (2X) and higher order quiz (2X) conditions.

In general, the fact quiz (2X) performance resulted in substantially greater performance during both the first and second quiz blocks compared with the other initial learning conditions, likely because of differences in item difficulty between fact and higher order questions. Post hoc comparisons confirmed that the fact quiz (2X) condition resulted in greater performance than the higher order quiz (1X) and higher order quiz (2X) conditions on the first quiz block, $t(47) = 4.00$, $p < .001$, $d = 0.71$ and $t(47) = 2.66$, $p = .011$, $d = 0.56$, respectively, but fact quiz (2X) performance was not significantly greater than mixed quiz (2X) performance on the first quiz block, $t(47) = 1.91$, $p > .05$, likely because the mixed quiz condition includes subjects whose first quiz was also a fact quiz. On the second quiz block, the fact quiz (2X) condition resulted in greater performance than the higher order quiz (2X) and mixed quiz (2X) conditions, $t(47) = 4.29$, $p < .001$, $d = 0.77$ and $t(47) = 15.66$, $p < .001$, $d = 3.13$, respectively.

Final test performance. Final test performance for the four retrieval practice conditions is displayed in Table 3 and Figure 3. Reliability (Cronbach's alpha) was .462 for final fact test performance and .254 for final higher order performance. There was no effect of counterbalancing order on final test performance for the mixed quiz (2X) condition (fact-higher order or higher order-fact), $F < 1$, therefore means were collapsed over counterbalance order for subsequent analyses (see Table 3 for the complete set of means).

As seen on the far right side of Table 3, delayed test performance was greatest for the mixed quiz (2X) condition (75%), compared with the fact quiz (2X, 69%), higher order quiz (2X, 69%), and higher order quiz (1X, 65%) conditions, respectively. Overall performance for the two test types was similar: 69% correct on the final fact test and 70% correct on the final higher order test. A 4 [retrieval practice condition: higher order quiz (1X), higher order quizzes (2X), fact quizzes (2X), mixed quizzes (2X)] × 2 (delayed test type: fact, higher order) repeated measures ANOVA on delayed test performance revealed a main effect of retrieval practice condition, $F(3, 141) = 4.85$, $p = .003$, $\eta_p^2 = .09$, and a significant interaction between retrieval practice condition and delayed test type, $F(3, 141) = 86.23$, $p < .001$, $\eta_p^2 = .65$.

Regarding delayed performance on the fact test, post hoc t tests confirmed that the fact quiz (2X) condition (90%) and the mixed

Table 3
Initial Quiz and Delayed Test Performance (Proportion Correct) as a Function of Retrieval Practice Condition From Experiment 2

Condition	First quiz	Second quiz	Final fact test	Final higher order test	Delayed average
Higher order quiz (1X)	.47 (.11)		.54 (.23)	.77 (.17)	.65
Higher order quizzes (2X)	.49 (.14)	.83 (.12)	.53 (.22)	.85 (.13)	.69
Fact quizzes (2X)	.57 (.17)	.91 (.08)	.90 (.13)	.48 (.19)	.69
Mixed quizzes (2X)	.52 (.19)	.53 (.15)	.78 (.18)	.71 (.18)	.75
Mixed: Fact-higher	.58 (.22)	.47 (.11)	.81 (.17)	.71 (.18)	.76
Mixed: Higher-fact	.45 (.13)	.60 (.15)	.76 (.18)	.71 (.19)	.73
Average	.53	.76	.69	.70	

Note. Standard deviations are displayed in parentheses.

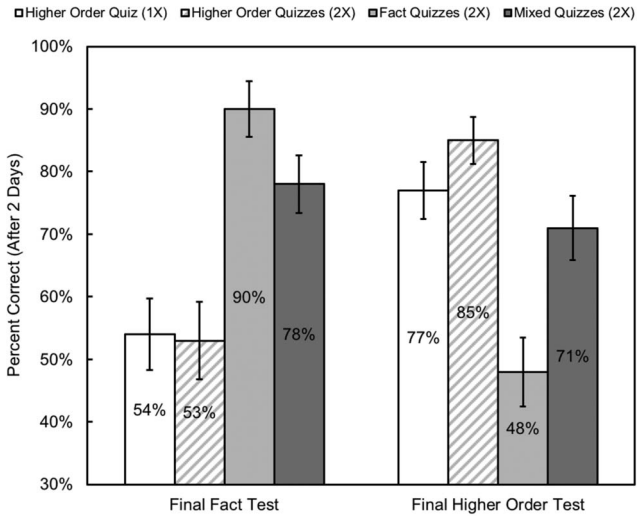

Figure 3. Delayed test performance (proportion correct after two days) as a function of retrieval practice condition from Experiment 2. Errors bars represent 95% confidence intervals.

quiz (2X) condition (78%) resulted in greater delayed test performance compared with the higher order quiz (1X, 54%) and the higher order quiz (2X, 53%) conditions, $ts > 6.10$, $ps < .001$, $ds > 1.21$. The difference in delayed fact test performance between the fact quiz (2X) and mixed quiz (2X) conditions was also significant, $t(47) = 3.72$, $p = .006$, $d = 0.77$.

Regarding delayed performance on the higher order test, post hoc t tests confirmed that the higher order quiz (2X, 85%), higher order quiz (1X, 77%), and mixed quiz (2X, 71%) conditions resulted in greater delayed test performance compared with the fact quiz (2X) condition (48%), $ts > 5.80$, $ps < .001$, $ds > 1.24$. The difference between the higher order quiz (2X) and the mixed quiz (2X) conditions was also significant, $t(47) = 4.52$, $p < .001$, $d = 0.84$; however, neither of these two conditions differed significantly from the higher order quiz (1X) condition, $ps > .05$.

Lastly, the difference in delayed performance between the congruent conditions, namely delayed fact test performance for the fact quiz (2X) condition (90%) and delayed higher order test performance for the higher order quiz (2X) condition (85%), was not significant, $t(47) = 2.01$, $p > .05$, and performance was close to ceiling levels. The difference between the mixed quiz (2X) condition on the delayed fact test (78%) versus the delayed higher order test (71%) was marginally significant, $t(47) = 2.08$, $p = .088$, $d = 0.39$.

Consistent with Experiment 1, the congruent conditions (fact quizzes-fact test, higher order quizzes-higher order test) resulted in the greatest delayed test performance compared with the mixed quiz (2X) condition, suggesting that two quizzes of the same format are more potent for long-term learning than one quiz of each format. Interestingly, the fact quiz (2X) condition still did not benefit delayed higher order performance, even when compared with only one initial higher order quiz, providing further evidence that a boost in fact learning does not necessarily improve delayed higher order performance.

Discussion

In Experiment 2, retrieval practice with two higher order quizzes improved delayed higher order test performance by an additional 8% compared with only one higher order quiz (see Figure 3). When the type of initial quizzes matched the type of final test, retrieval practice yielded benefits for both fact and higher order learning to a greater extent than one quiz of each format (in the mixed quiz condition). Replicating Experiment 1, retrieval practice with fact questions did not enhance delayed higher order test performance, inconsistent with the foundation of factual knowledge framework. In addition, retrieval practice with higher order questions did not enhanced delayed fact test performance. The findings from Experiment 2 provide further evidence that retrieval practice is the most powerful when questions encountered during initial quizzes are similar to questions on a final test, consistent with the transfer appropriate processing framework.

Experiment 3

Experiment 3 was designed to investigate whether results from Experiments 1 and 2 would replicate in an applied setting with a different population (6th grade students) and a different content domain (world history). In previous research with middle school students, retrieval practice enhanced delayed test performance compared with no quizzes, although information to be learned was mainly fact-based (Agarwal et al., 2012; McDaniel, Agarwal, Huelser, McDermott, & Roediger, 2011; McDaniel et al., 2013; McDermott et al., 2014; Roediger et al., 2011). Do younger students receive the same benefit from mixed quizzing as do college students? It may be the case that, particularly for younger students, the close succession of both question types within one quiz facilitates students' building of connections across the two types of processing, while also providing a prompt or hint to transfer knowledge from fact questions to the immediate higher order questions (Gick & Holyoak, 1980; Pan & Rickard, in press).

In Experiment 3, students completed mixed quizzes or higher order-only quizzes, and learning was measured on final fact and higher order tests. In accordance with prior research, retrieval practice (regardless of quiz condition) was expected to enhance both delayed fact and higher order test performance, compared with delayed test performance on nonquizzed items. Based on findings from Experiments 1 and 2, higher order quizzes were expected to enhance delayed higher order test performance, but not delayed fact test performance. Regarding mixed quizzes, the transfer appropriate processing framework suggests a partial improvement for both types of delayed tests, based on a partial overlap of similar question types (see Table 1).

Method

Participants. One hundred forty-two 6th grade students ($M = 24$ students in each of six classroom sections; 71 males, 71 females) from a Midwestern suburban middle school participated in the current experiment. Assent from each student was obtained in accordance with guidelines from the Human Research Protection Office. Twelve students declined to include their data in the study (although they participated in all classroom activities), and data from eight special education students were excluded from

analyses because they received accommodations (e.g., additional study and quiz opportunities outside of class).

Design. A 3 × 2 within-subjects design was used, such that three retrieval practice conditions (higher order quizzes, mixed quizzes, nonquizzed) were crossed with two delayed test types (fact test, higher order test). Conditions were manipulated across two chapters of Social Studies material, with chapters presented in the same order for all students (as determined by the classroom teacher). Six classroom sections were split into two sets of three class sections each; in other words, sections 1, 3, and 6 constituted Set A, and sections 2, 5, and 7 constituted Set B. For the first chapter, Set A students completed three quizzes with higher order questions, whereas Set B students completed three quizzes with a mix of question types. For the second chapter, the retrieval practice conditions switched. At the end of each chapter unit (approximately 7–8 school days in length; 48 hr after the third quiz), students completed a final test comprised of both question types (fact and higher order), with all questions presented in a different random order for each of the six classroom sections. To maximize power using the largest number of items per condition as possible, while reducing classroom time required for the manipulation, a restudy condition was not included in this experiment (prior research demonstrated that retrieval practice enhanced delayed test performance compared with a restudy exposure control in the same school; McDermott et al., 2014; Roediger et al., 2011).

Materials. Two social studies textbook chapters (Russian Revolution and World War II; Banks et al., 1997), assigned by the classroom teacher, were used in this experiment. Each chapter was approximately 2,350 words in length (e.g., 2,335 words for the Russian Revolution chapter and 2,407 words for the World War II chapter).

Twelve four-alternative multiple-choice fact questions and 12 four-alternative multiple-choice higher order questions were developed for each textbook chapter (see Appendix B for sample questions). The classroom teacher approved all questions and multiple-choice alternatives. Across all initial quizzes and delayed tests, each classroom section received a different set of quizzed and nonquizzed items, and every item was quizzed or not quizzed at least once. In addition, for every initial quiz and final test, each classroom section received a different random order of questions and the multiple-choice alternatives were randomly reordered. The correct multiple-choice alternative appeared in every position (A, B, C, or D) an equal number of times across quizzes and tests.

Initial quizzes comprised either eight higher order questions or a mix of fact and higher order questions (four of each type). For mixed quizzes, question type (fact or higher order) was blocked and order was counterbalanced across classroom sections and quizzes, with questions presented in random order within question type block. Questions that were not on initial quizzes (a nonquizzed control condition) were covered in the textbook chapter and also during the teacher's lessons.

Final chapter tests comprised all multiple-choice fact and higher order questions (12 fact and 12 higher order questions per chapter). Final chapter test questions and alternatives were the same as those from initial quizzes (i.e., questions were not rephrased) because of the teacher's concerns regarding floor effects and all 24 items were presented in random order (not blocked by question type) on final chapter tests. All reading passages and questions developed are included in the online supplementary material.

For fact questions, broad concepts stated in the chapters were tested to measure students' overall understanding of the content. For example, a fact question from the "Russian Revolution" textbook chapter included:

> Why was Nicholas II forced to give up his role as tsar?
>
> A. Because the Duma elected a new tsar
> B. Because Stalin took over the government
> C. Because his wife and children moved to Moscow
> D. Because of angry protestors, soldiers, and railroad workers

The correct answer for this fact question is alternative D and this answer was stated directly in the textbook chapter. In contrast to typical laboratory experiments, all fact questions in the present study were drawn from authentic classroom material and designed to encompass key concepts or ideas from the textbook chapters, rather than details such as names, dates, vocabulary words, definitions, and so forth (e.g., the year in which the Russian Revolution began).

Higher order questions were developed based on categories from a revised Bloom's taxonomy (*apply*, *analyze,* and *evaluate;* Anderson et al., 2001; see Figure 1). For example, an *analyze* question from the "Russian Revolution" chapter included:

> Which person would agree with the following statement? "Revolutions are hard to prevent."
>
> A. Alexander II
> B. Lenin
> C. Nicholas II
> D. Stalin

The correct answer for this *analyze* question is alternative C. Higher order questions from the taxonomic *create* category were not included in this experiment, due to the teacher's concerns that 6th grade students may have difficulty extending textbook facts to completely novel situations (i.e., floor effects).

Procedure. Students completed initial quizzes individually via a clicker response system (Ward, 2007) in the classroom using a computer, projector, and projection screen at the front of the classroom. At the beginning of the study, students were instructed that they would be taking quizzes (via clickers, with which students were already familiar) and tests as part of a research study, and that their scores may or may not count for a grade. In actuality, students' scores were not factored into their individual grades; instead, the average score for each of the six classroom sections counted toward a pizza party held at the end of the school year. The classroom section with the highest score on each initial quiz or final test received five points toward the pizza party. The classroom section with the second highest score on each quiz or test received four points toward the pizza party. Additional classroom assignments and exams also factored into students' pizza party point totals, as determined by the classroom teacher.

In general, each chapter unit lasted approximately one week. For each unit, students read a chapter from their social studies textbook, listened to seven or eight lessons, participated in quizzes (the experimental manipulation), and completed standard assignments developed by the teacher.

Before the first lesson, students completed a prequiz via clickers without having read the textbook chapter. Immediately after the prequiz (i.e., on the same day), students began reading the chapter and participated in the teacher's corresponding lesson. After 2–3 school days, during which students completed the chapter reading and the teacher covered all chapter content, students completed a postquiz via clickers. Two days after the postquiz, the classroom teacher reviewed all chapter material, which was followed by a review quiz via clickers.

During all clicker quizzes (pre-, post-, and review quizzes), multiple-choice questions were displayed on a projection screen at the front of the classroom one at a time, in a different random order for each classroom section. I read the question stem and four multiple-choice alternatives aloud. After I was finished reading the question and alternatives, I made the software's response option available and students were asked to press a letter (A, B, C, or D) on their clicker remote corresponding to the multiple-choice alternative (forced choice). Once all students in the classroom responded (after approximately one minute), I closed the response option and the clicker software displayed a green checkmark next to the correct alternative (i.e., immediate feedback was administered during quizzes). I read aloud the question stem and the correct answer, and then moved on to the next question. Each clicker quiz comprised eight questions, which took each class section approximately seven to nine minutes to complete. Mental effort ratings were not collected.

Two days after the review quiz, students completed a final test comprised of all 24 multiple-choice questions for the chapter. Final chapter tests were administered online (via Google Docs, http://docs.google.com), whereas students sat individually at PC computers. The chapter test was self-paced and students viewed each multiple-choice question one at a time. Once students selected a multiple-choice alternative, they moved on to the next question; however, the online chapter test also allowed students to return to earlier questions if they wanted to review or change their answers. Once students responded to all 24 test questions, students were no longer able to return to the test to change their answers. No feedback was provided during the final chapter test.

In sum, students participated in three within-subject retrieval practice conditions, crossed with two final test types. The dependent variable of interest was accuracy on final test questions. The entire procedure was followed for two textbook chapters over the course of two weeks. At the end of the study, students were debriefed and thanked for their time.

Results

Thirty-four students were absent during at least one initial quiz or the final chapter test and their data were excluded from the reported analyses to ensure the integrity of the experimental manipulation. Thus, data reported are from 88 students (M age = 11.58 years, 48 females); similar patterns of results were found when data from all students who assented to participate were included (i.e., n = 122 absent and present students, excluding special education students).

Note that middle school students were assigned to class sections before the current study began and this assignment was nonrandom. Thus, data from students are nested within class section, and again nested within set. Experiment 3 was carried out completely within-subjects, all materials and conditions were completely counterbalanced, and there was no significant difference in final test performance between the two sets (p = .082). Even so, nested individuals tend to be more alike than individuals selected at random (Raudenbush & Bryk, 2002). Because the number of levels within nests was low, the use of a multilevel model to determine the influence of nonrandom assignment on performance was not possible and means have not been adjusted to account for this nesting.

Initial quiz performance. Initial quiz performance for the first (prequiz), second (postquiz), and third (review) quizzes is displayed in Table 4. In general, initial performance increased from the prequiz (38%) to the postquiz (71%) and also to the review quiz (84%), as a result of textbook reading, classroom lessons, and immediate feedback received during the clicker quizzes. Across the initial quizzes, performance was slightly greater in the mixed quiz condition (66%) compared with performance in the higher order quiz condition (62%), likely because of the inclusion of fact questions in the mixed quiz condition.

A 2 (retrieval practice condition: higher order quizzes, mixed quizzes) × 3 (quiz type: pre, post, review) repeated measures ANOVA on initial quiz performance revealed a marginal main effect of retrieval practice condition, $F(1, 87) = 3.55$, $p = .063$, $\eta_p^2 = .039$, and a significant main effect of quiz type, $F(2, 174) = 442.05$, $p < .001$, $\eta_p^2 = .84$; however, the interaction was not significant, $F(2, 174) = 2.03$, $p > .05$. In other words, students' initial quiz performance increased across the three quizzes, and did so similarly for the mixed quiz and the higher order quiz conditions.

Final test performance. Performance on the final tests, administered two days after the review quizzes, is displayed in Table 4 and Figure 4. Reliability (Cronbach's alpha) was .709 for final fact test performance and .686 for final higher order performance. A univariate ANOVA with set as a between-subjects factor (two sets of three classroom sections each; see Appendix A) revealed no significant difference on overall final test performance, $F(1, 87) = 3.12$, $p = .082$, therefore means were collapsed over set for subsequent analyses.

Table 4
Initial Quiz and Delayed Test Performance (Proportion Correct) as a Function of Retrieval Practice Condition From Experiment 3

Condition	Pre-quiz	Post-quiz	Review quiz	Final fact test	Final higher order test	Delayed average
Non-quizzed				.64 (.18)	.56 (.18)	.60
Higher order quizzes	.38 (.16)	.68 (.21)	.82 (.17)	.64 (.20)	.75 (.21)	.70
Mixed quizzes	.38 (.18)	.73 (.19)	.87 (.15)	.91 (.17)	.82 (.21)	.86
Average	.38	.71	.84	.73	.71	

Note. Standard deviations are displayed in parentheses.

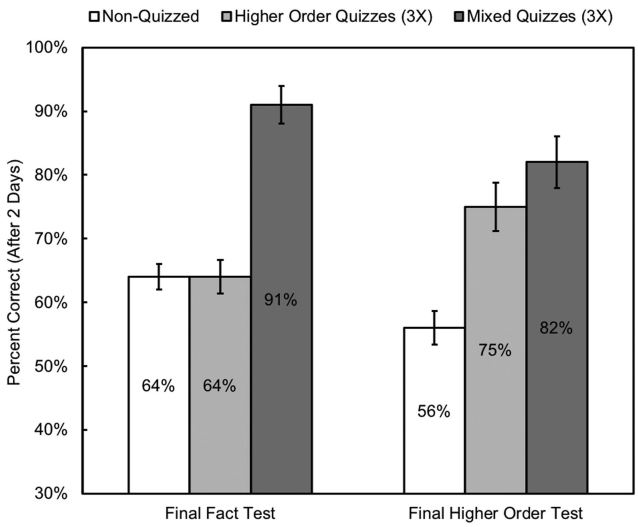

Figure 4. Delayed test performance (proportion correct after two days) as a function of retrieval practice condition from Experiment 3. Errors bars represent 95% confidence intervals.

Delayed test performance (collapsed over test type) was greatest for the mixed quiz condition (86%), followed by the higher order quiz (70%) and nonquizzed (60%) conditions. In addition, delayed performance was similar for the final fact (73%) and final higher order (71%) tests. A 3 (retrieval practice condition: higher order quizzes, mixed quizzes, nonquizzed) × 2 (final test type: fact, higher order) repeated measures ANOVA on final test performance revealed a significant main effect of retrieval practice condition, $F(2, 174) = 128.98$, $p < .001$, $\eta_p^2 = .60$, a marginal main effect of final test type, $F(1, 87) = 3.19$, $p = .078$, $\eta_p^2 = .04$, and a significant interaction between retrieval practice condition and final test type, $F(2, 174) = 28.30$, $p < .001$, $\eta_p^2 = .25$.

For the final fact test, the mixed quiz condition resulted in far greater performance (91%) than the higher order quiz and nonquizzed conditions (64% each), $t(47) = 12.24$, $p < .001$, $d = 1.44$ and $t(47) = 13.63$, $p < .001$, $d = 1.55$, respectively. For the final higher order test, the mixed quiz condition again resulted in the greatest performance (82%) compared with the higher order (75%) and nonquizzed (56%) conditions, $t(87) = 2.27$, $p = .078$ ($p = .026$ without Bonferroni correction), $d = 0.34$ and $t(87) = 12.24$, $p < .001$, $d = 1.37$, respectively. Consistent with Experiments 1 and 2, higher order retrieval practice led to significantly greater higher order test performance compared with the nonquizzed condition, $t(87) = 7.87$, $p < .001$, $d = 0.99$.

Overall, mixed retrieval practice produced the greatest level of performance on both fact and higher order final tests, while providing a marginal benefit above and beyond the benefit from higher order retrieval practice on delayed higher order test performance.

Discussion

Retrieval practice dramatically increased learning for middle school students compared with no quizzes, contributing to a growing body of applied research (McDaniel et al., 2011, 2013; McDermott et al., 2014; Roediger et al., 2011). Remarkably, mixed quizzes increased final fact test performance from the letter grade equivalent of a D to an A− (a difference of 27%, Figure 4). Mixed quizzes also produced a slight improvement over and above higher order quizzes on delayed higher order test performance (7%), although this difference was marginally significant. Replicating Experiments 1 and 2, findings were consistent with the transfer appropriate processing framework (see Table 1): benefits from higher order quizzes were limited to the higher order test, and benefits from the mixed quizzes extended to both types of final tests.

General Discussion

Contrary to popular intuition, building a foundation of factual knowledge via retrieval practice did not enhance students' higher order learning. Instead, students' final fact test and higher order test performance was greatest following retrieval practice that matched in cognitive complexity based on Bloom's taxonomy: fact quizzes enhanced final fact test performance and higher order quizzes enhanced final higher order test performance. Retrieval practice increased learning by 20–30% under laboratory conditions with college students and also in an authentic K-12 classroom.

Fact Quizzes Do Not Enhance Higher Order Learning

Why didn't fact quizzes improve higher order learning in the present study, as many cognitive scientists and educators contend? First, students may have been unaware that information on fact quizzes was related to final higher order tests, thus they did not transfer their knowledge without explicit instructions to do so. Chan, McDermott, and Roediger (2006, Experiment 3) found a benefit of retrieval practice on novel items when subjects were instructed to adopt a "broad retrieval strategy" during study, whereas subjects who were told to adopt a "narrow retrieval strategy" did not demonstrate a benefit of retrieval practice on related novel items. Butler (2010, Experiment 3) also found a benefit of retrieval practice on far transfer to novel items when subjects were explicitly told that the final test was related to information learned during the initial session (see also Chan, 2009). Furthermore, a classic study by Gick and Holyoak (1980) demonstrated that students' conceptual knowledge remains "inert" when not explicitly told to use previously learned information on novel items (see also Bransford, Sherwood, Vye, & Rieser, 1986; Pan & Rickard, in press). Thus, students in the current study may have transferred their factual knowledge to the higher order test questions had they been given explicit instructions, prompts, or hints.

Second, quiz and final test questions were multiple-choice, a retrieval practice format that improves learning in laboratory and classroom settings (Bjork, Little, & Storm, 2014; McDermott et al., 2014; Roediger, Agarwal, Kang, & Marsh, 2010). Some educators argue that higher order learning cannot be facilitated or measured using multiple-choice quizzes or tests. Instead, educators often advocate for paper assignments, essay tests, open-book tests, and ongoing portfolio evaluations to determine higher order learning (Ausubel et al., 1978; Hart, 1994; Kohn, 1999; Martinez, 1999). It is possible that by using multiple-choice items, an element of genuine higher order learning may have been lost in the

present study. Alternatively, multiple-choice quizzes may have provided an *added* benefit from short answer quizzes because well-constructed multiple-choice alternatives can facilitate retrieval of information pertaining to correct and incorrect alternatives, as well as enhance transfer to novel information (Gierl, Bulut, Guo, & Zhang, 2017; Little & Bjork, 2015; Little, Bjork, Bjork, & Angello, 2012; Marsh & Cantor, 2014).

In the present study, multiple-choice alternatives were developed to require students to make fine-grained distinctions between similar concepts. For instance, in Experiments 1 and 2, college students were asked to evaluate an author's views on welfare and whether the government's primary role is to advance morality, security, equality, or liberty (see Appendix B). Each of these multiple-choice alternatives related to the reading passage, the target question, and current policy on welfare programs. As such, the multiple-choice alternatives may have *increased*, rather than decreased, potential transfer from fact to higher order questions. Further examination of higher order learning with a variety of question formats—including multiple-choice—may shed light on whether students' transfer of fact knowledge to higher order learning is dependent upon the type of questions used during retrieval practice.

It is important to note that all multiple-choice alternatives remained identical from initial retrieval practice to final tests for all three experiments. In addition, Experiments 1 and 2 included slightly rephrased final test questions, whereas Experiment 3 included final test questions that were identical to initial quiz questions. Recent research has found that rephrased quiz or test questions do not facilitate transfer to a greater extent than identical questions (Butler, 2010; Pan & Rickard, in press); thus, although rephrased questions would be ideal for examining the flexibility of learning, it is uncertain whether materials in the current study diminished or eliminated transfer of fact knowledge. Further, if memorization of quiz and test items were a concern, then one would expect ceiling performance greater than 90% on final tests, which was not found in the present experiments.

Third, cognitive load theory suggests that fact quizzes should facilitate higher order learning by reducing cognitive demands required during the final test (Plass et al., 2010; Sweller, 2010; van Gog & Sweller, 2015). It was expected that mental effort ratings on higher order tests would be lower when preceded by fact quizzes compared with higher order quizzes (rated on a 9-point scale, adapted from Paas, 1992; data available upon request). Unfortunately, mental effort ratings did not shed light on this puzzle. Across experiments and conditions, mental effort ratings during higher order tests were lower when students first completed higher order quizzes. In other words, students' mental effort ratings did not reveal sensitivity to a foundation of factual knowledge on higher order learning.

The key finding that fact quizzes did not enhance higher order learning directly contradicts the long-held belief that "factual knowledge must precede skill" (Willingham, 2009). Instead, a match between initial quiz and final test questions produced the greatest learning in the current study, consistent with the transfer appropriate processing framework (Morris et al., 1977; see Table 1, rows 1 and 2). In addition, discrepant conditions in the present study (fact quiz-higher order test and higher order quiz-fact test; see Table 1, rows 3 and 4) did not promote learning, contrary to recent findings (Hinze & Wiley, 2011; Jensen et al., 2014; McDermott et al., 2014). Whether a foundation of factual knowledge promotes higher order learning—and under what conditions—remains to be seen. Exploration of prompts to transfer and question format may yield fact-based retrieval strategies that push student learning higher on Bloom's taxonomy and to greater levels of complexity.

Mixed Quizzes Enhance Higher Order Learning

Mixed quizzes, comprising both fact and higher order questions, increased higher order test performance more than fact quizzes (in Experiment 2) and slightly more than higher order quizzes (in Experiment 3). The robust benefits of mixed quizzes on higher order learning is consistent with predictions from the three theoretical frameworks of interest: the desirable difficulty framework suggests that mixed quizzes pose a challenge for students by switching between question complexity during retrieval (see also Butler et al., 2017); the transfer appropriate processing framework suggests a benefit of mixed quizzes because of an overlap in processing with the final test; and the foundation of factual knowledge framework predicts a benefit from mixed quizzes when students are given the opportunity to engage in fact questions during initial learning (see Table 1, rows 5 and 6).

Compared with higher order quizzes, why were mixed quizzes more potent for higher order learning in Experiment 3 (middle school students) than in Experiment 2 (college students)? Of course, students of different ages may benefit differentially from mixed retrieval practice. Varied complexity may be advantageous by providing a scaffold between lower order and higher order questions, particularly for students who have limited experience with complex materials (i.e., 6th grade students in Experiment 3). Meanwhile, scaffolding from varied complexity may be unnecessary for students who already have experience extracting factual information from complex materials (i.e., college students in Experiment 2); thus, older students may reap more benefits from higher order retrieval practice. Although pure speculation, the current study is the first to implement a retrieval practice paradigm under similar conditions for two distinct student populations. More research is necessary to ascertain whether factual knowledge improves higher order learning, but also to ascertain whether it differs for different age groups. Support for the foundation of factual knowledge framework is frequently articulated from a K-12 perspective, an indication that the relationship between fact learning and higher order learning might be unique for younger students.

Another major difference between Experiment 2 and Experiment 3 was the procedure followed for mixed quizzes. In Experiment 2, mixed quizzes were counterbalanced across numerous within-subject conditions (e.g., students completed a fact quiz on the welfare passage, next completed quizzes on other passages, and then completed a higher order quiz on the welfare passage; see Appendix A). In contrast, mixed quizzes in Experiment 3 comprised fact and higher order questions concurrently. Furthermore, within-subject conditions were administered one after another during a single session in Experiment 2, whereas middle school students in Experiment 3 spent one week with one type of retrieval practice (higher order or mixed) and then spent another week with the other type of retrieval practice.

These procedural differences may have influenced students' motivation to pay attention and learn the material, resulting in

different benefits from mixed quizzing in the two experiments. Prior research indicates that students prefer learning activities that vary in terms of item difficulty, particularly in mathematics (Skinner, Fletcher, Wildmon, & Belfiore, 1996; Wildmon, Skinner, & McDade, 1998). Recent research also indicates that students adjust their learning strategies based on expectancy for different types of tests (Agarwal, D'Antonio, Roediger, McDermott, & McDaniel, 2014; Agarwal & Roediger, 2011; Jensen et al., 2014). Broadly, it would have been challenging for college students to modulate their motivation specifically for the mixed condition because of the extensive counterbalancing of passages and quizzes. It is possible that middle school students exerted more motivational effort during the week in which they had mixed quizzes compared with the week of higher order quizzes, even though students were provided with an incentive (a pizza party) based on quiz and test performance during both weeks.

It is also unlikely that students in Experiments 2 and 3 would have had greater motivation in the mixed quiz condition due to the novelty of quiz items. To recognize which conditions were mixed or not mixed, college students would have needed to keep track of multiple-choice items across seven intervening quizzes on unrelated passages, and middle school students would have needed to ascertain the distinction between a change in chapter material versus a change in the complexity of quiz questions. Considering the amount of attention required to do this in either experiment, the likelihood that the mixed quiz condition increased student motivation as a result of item novelty is doubtful. (I thank an anonymous reviewer for bringing this possibility to my attention.)

When it comes to instruction, what type of retrieval practice will help students achieve the highest levels of Bloom's taxonomy? Surprisingly, fact-based retrieval practice only increased fact learning, whereas higher order and mixed retrieval practice increased higher order learning. If we want to reach the top of Bloom's taxonomy, building a foundation of knowledge via fact-based retrieval practice may be less potent than engaging in higher order retrieval practice at the outset, a key finding for future research and classroom application.

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,* 659–701. http://dx.doi.org/10.3102/0034654316689306

Agarwal, P. K., Bain, P. M., & Chamberlain, R. W. (2012). The value of applied research: Retrieval practice improves classroom learning and recommendations from a teacher, a principal, and a scientist. *Educational Psychology Review, 24,* 437–448. http://dx.doi.org/10.1007/s10648-012-9210-2

Agarwal, P. K., D'Antonio, L., Roediger, H. L., III, McDermott, K. B., & McDaniel, M. A. (2014). Classroom-based programs of retrieval practice reduce middle school and high school students' test anxiety. *Journal of Applied Research in Memory & Cognition, 3,* 131–139. http://dx.doi.org/10.1016/j.jarmac.2014.07.002

Agarwal, P. K., Finley, J. R., Rose, N. S., & Roediger, H. L., III. (2017). Benefits from retrieval practice are greater for students with lower working memory capacity. *Memory, 25,* 764–771. http://dx.doi.org/10.1080/09658211.2016.1220579

Agarwal, P. K., Karpicke, J. D., Kang, S. H. K., Roediger, H. L., III, & McDermott, K. B. (2008). Examining the testing effect with open- and closed-book tests. *Applied Cognitive Psychology, 22,* 861–876. http://dx.doi.org/10.1002/acp.1391

Agarwal, P. K., & Roediger, H. L., III. (2011). Expectancy of an open-book test decreases performance on a delayed closed-book test. *Memory, 19,* 836–852. http://dx.doi.org/10.1080/09658211.2011.613840

Agarwal, P. K., Roediger, H. L., McDaniel, M. A., & McDermott, K. B. (2017). *How to use retrieval practice to improve learning*. St. Louis, MO: Washington University in St. Louis. Retrieved from http://www.retrievalpractice.org

Anderson, L. W., Krathwohl, D. R., Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R., . . . Wittrock, M. C. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives* (abridged ed.). New York, NY: Addison Wesley Longman.

Ausubel, D. P. (1965). In defense of verbal learning. In R. Anderson & D. Ausubel (Eds.), *Readings in the psychology of cognition* (pp. 87–102). New York, NY: Holt, Rinehart, & Winston. (Original work published 1961)

Ausubel, D. P., Novak, J. D., & Hanesian, H. (1978). *Educational psychology: A cognitive view* (2nd ed.). New York, NY: Holt, Rinehart, and Winston.

Banks, J. A., Beyer, B. K., Contreras, G., Craven, J., Ladson-Billings, G., McFarland, M. A., & Parker, W. C. (1997). *World: Adventures in time and place*. New York, NY: Macmillan/McGraw-Hill.

Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn? A taxonomy for far transfer. *Psychological Bulletin, 128,* 612–637. http://dx.doi.org/10.1037/0033-2909.128.4.612

Bartlett, F. C. (1958). *Thinking: An experimental and social study*. Westport, CT: Greenwood Press.

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 & Cognition, 3,* 165–170. http://dx.doi.org/10.1016/j.jarmac.2014.03.002

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

Bloom, B. S. (Ed.), Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *The taxonomy of educational objectives: The classification of educational goals* (Handbook 1: Cognitive domain). New York, NY: David McKay Company.

Bransford, J. D., Sherwood, R., Vye, N., & Rieser, J. (1986). Teaching thinking and problem solving: Research foundations. *American Psychologist, 41,* 1078–1089. http://dx.doi.org/10.1037/0003-066X.41.10.1078

Brown, P. C., Roediger, H. L., & McDaniel, M. A. (2014). *Make it stick: The science of successful learning*. Cambridge, MA: Harvard University Press. http://dx.doi.org/10.4159/9780674419377

Bruner, J. S. (1977). *The process of education*. Cambridge, MA: Harvard University Press.

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

Butler, A. C., 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,* 433–446. http://dx.doi.org/10.1037/xap0000142

Butler, A. C., Marsh, E. J., Slavinsky, J. P., & Baraniuk, R. G. (2014). Integrating cognitive science and technology improves learning in a STEM classroom. *Educational Psychology Review, 26,* 331–340. http://dx.doi.org/10.1007/s10648-014-9256-4

Callender, A. A., & McDaniel, M. A. (2009). The limited benefits of rereading educational texts. *Contemporary Educational Psychology, 34,* 30–41. http://dx.doi.org/10.1016/j.cedpsych.2008.07.001

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

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

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

Chan, J. C. K. (2009). When does retrieval induce forgetting and when does it induce facilitation? Implications for retrieval inhibition, testing effect, and text processing. *Journal of Memory and Language, 61,* 153–170. http://dx.doi.org/10.1016/j.jml.2009.04.004

Chan, J. C. K., McDermott, K. B., & Roediger, H. L., III. (2006). Retrieval-induced facilitation: Initially nontested material can benefit from prior testing of related material. *Journal of Experimental Psychology: General, 135,* 553–571. http://dx.doi.org/10.1037/0096-3445.135.4.553

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cousineau, D. (2005). Confidence intervals in within-subject designs: A simpler solution to Loftus and Masson's method. *Tutorials in Quantitative Methods for Psychology, 1,* 42–45. http://dx.doi.org/10.20982/tqmp.01.1.p042

Cuban, L. (1984). Policy and research dilemmas in the teaching of reasoning: Unplanned designs. *Review of Educational Research, 54,* 655–681. http://dx.doi.org/10.3102/00346543054004655

Daniel, E. L. (Ed.), (2006). *Taking sides: Clashing views in health and society* (7th ed.). Dubuque, IA: McGraw-Hill Companies, Inc.

Dewey, J. (1944). *Democracy and education: An introduction to the philosophy of education*. New York, NY: The Free Press. (Original work published 1916)

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

Easton, T. A. (Ed.), (2006). *Taking sides: Clashing views on environmental issues* (11th ed.). Dubuque, IA: McGraw-Hill Companies, Inc.

Finsterbusch, K., & McKenna, G. (Eds.). (1984). *Taking sides: Clashing views on controversial social issues* (3rd ed.). Guilford, CT: Dushkin Publishing Group.

Gardiner, F. M., Craik, F. I. M., & Bleasdale, F. A. (1973). Retrieval difficulty and subsequent recall. *Memory & Cognition, 1,* 213–216. http://dx.doi.org/10.3758/BF03198098

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

Gierl, M. J., Bulut, O., Guo, Q., & Zhang, X. (2017). Developing, analyzing, and using distractors for multiple-choice tests in education: A comprehensive review. *Review of Educational Research, 87,* 1082–1116. http://dx.doi.org/10.3102/0034654317726529

Greenhouse, S. W., & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika, 24,* 95–112. http://dx.doi.org/10.1007/BF02289823

Hart, D. (1994). *Authentic assessment: A handbook for educators*. Menlo Park, CA: Addison Wesley Publishing Company.

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

Hirsch, E. D. (1996). *The schools we need and why we don't have them*. New York, NY: Doubleday.

Jacoby, L. L., Wahlheim, C. N., & Coane, J. H. (2010). Test-enhanced learning of natural concepts: Effects on recognition memory, classification, and metacognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36,* 1441–1451. http://dx.doi.org/10.1037/a0020636

James, W. (1900). *Talks to teachers on psychology: And to students on some of life's ideals*. New York, NY: Henry Holt and Company.

Jensen, J. L., McDaniel, M. A., Woodard, S. M., & Kummer, T. A. (2014). Teaching to the test . . . or testing to teach: Exams requiring higher order thinking skills encourage greater conceptual understanding. *Educational Psychology Review, 26,* 307–329. http://dx.doi.org/10.1007/s10648-013-9248-9

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

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

Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science, 331,* 772–775. http://dx.doi.org/10.1126/science.1199327

Karpicke, J. D., & Roediger, H. L., III. (2007). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language, 57,* 151–162. http://dx.doi.org/10.1016/j.jml.2006.09.004

Knapp, M. (2016, October 11). *5 gorgeous depictions of Bloom's taxonomy* [Blog post]. Retrieved from https://news.nnlm.gov/nto/2016/10/11/5-gorgeous-depictions-of-blooms-taxonomy/

Kohn, A. (1999). *The schools our children deserve: Moving beyond traditional classrooms and "tougher standards."* Boston, MA: Houghton Mifflin Company.

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

Kromann, C. B., Jensen, M. L., & Ringsted, C. (2009). The effect of testing on skills learning. *Medical Education, 43,* 21–27. http://dx.doi.org/10.1111/j.1365-2923.2008.03245.x

Larsen, D. P., Butler, A. C., Lawson, A. L., & Roediger, H. L., III. (2013). The importance of seeing the patient: Test-enhanced learning with standardized patients and written tests improves clinical application of knowledge. *Advances in Health Sciences Education, 18,* 409–425. http://dx.doi.org/10.1007/s10459-012-9379-7

Lemov, D. (2017, April 3). *Bloom's taxonomy: That pyramid is a problem* [Blog post]. Retrieved from http://teachlikeachampion.com/blog/blooms-taxonomy-pyramid-problem/

Little, J. L., & Bjork, E. L. (2015). Optimizing multiple-choice tests as tools for learning. *Memory & Cognition, 43,* 14–26. http://dx.doi.org/10.3758/s13421-014-0452-8

Little, J. L., Bjork, E. L., Bjork, R. A., & Angello, G. (2012). Multiple-choice tests exonerated, at least of some charges: Fostering test-induced learning and avoiding test-induced forgetting. *Psychological Science, 23,* 1337–1344. http://dx.doi.org/10.1177/0956797612443370

Lyle, K. B., & Crawford, N. A. (2011). Retrieving essential material at the end of lectures improves performance on statistics exams. *Teaching of Psychology, 38,* 94–97. http://dx.doi.org/10.1177/0098628311401587

Madaras, L., & SoRelle, J. M. (Eds.). (1993). *Taking sides: Clashing views on controversial issues in American history* (5th ed., Vol. 1). Guilford, CT: Dushkin Publishing Group.

Marsh, E. J., & Cantor, A. D. (2014). Learning from the test: Dos and don'ts for using multiple-choice tests. In M. A. McDaniel, R. F. Frey, S. M. Fitzpatrick, & H. L. Roediger (Eds.), *Integrating cognitive science*

with innovative teaching in STEM disciplines (pp. 37–52). St. Louis, MO: Washington University Libraries.

Martinez, M. E. (1999). Cognition and the question of test item format. *Educational Psychologist, 34,* 207–218. http://dx.doi.org/10.1207/s15326985ep3404_2

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

McDaniel, M. A., Friedman, A., & Bourne, L. E. (1978). Remembering the levels of information in words. *Memory & Cognition, 6,* 156–164. http://dx.doi.org/10.3758/BF03197441

McDaniel, M. A., Roediger, H. L., III, & McDermott, K. B. (2007). Generalizing test-enhanced learning from the laboratory to the classroom. *Psychonomic Bulletin & Review, 14,* 200–206. http://dx.doi.org/10.3758/BF03194052

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,* 360–372. http://dx.doi.org/10.1002/acp.2914

McDermott, K. B., Agarwal, P. K., D'Antonio, L., Roediger, H. L., III, & McDaniel, M. A. (2014). Both multiple-choice and short-answer quizzes enhance later exam performance in middle and high school classes. *Journal of Experimental Psychology: Applied, 20,* 3–21. http://dx.doi.org/10.1037/xap0000004

Mehta, J. (2018, January 4). *A pernicious myth: Basics before deeper learning* [Blog post]. Retrieved from http://blogs.edweek.org/edweek/learning_deeply/2018/01/a_pernicious_myth_basics_before_deeper_learning.html

Morey, R. D. (2008). Confidence intervals from normalized data: A correction to Cousineau (2005). *Tutorials in Quantitative Methods for Psychology, 4,* 61–64. http://dx.doi.org/10.20982/tqmp.04.2.p061

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer-appropriate processing. *Journal of Verbal Learning & Verbal Behavior, 16,* 519–533. http://dx.doi.org/10.1016/S0022-5371(77)80016-9

Moseley, W. G. (Ed.), (2007). *Taking sides: Clashing views on African issues* (2nd ed.). Dubuque, IA: McGraw-Hill Companies, Inc.

Münsterberg, H. (1909). *Psychology and the teacher*. New York, NY: D. Appleton and Company.

National Research Council. (1987). *Education and learning to think*. Washington, DC: The National Academies Press.

Noll, J. W. (Ed.), (2001). *Taking sides: Clashing views on controversial educational issues* (11th ed.). Guilford, CT: Dushkin/McGraw-Hill.

Paas, F. G. W. C. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. *Journal of Educational Psychology, 84,* 429–434. http://dx.doi.org/10.1037/0022-0663.84.4.429

Pan, S. C., & Agarwal, P. K. (2018). *Retrieval practice and transfer of learning: Fostering students' application of knowledge*. San Diego, CA: University of California at San Diego. Retrieved from http://www.retrievalpractice.org

Pan, S. C., & Rickard, T. C. (in press). Transfer of test-enhanced learning: Meta-analytic review and synthesis. *Psychological Bulletin*.

Pashler, H., Bain, P. M., Bottge, B. A., Graesser, A., Koedinger, K., McDaniel, M., & Metcalfe, J. (2007). *Organizing instruction and study to improve student learning*. Washington, DC: National Center for Education Research, Institute of Education Sciences, U.S. Department of Education. Retrieved from http://ncer.ed.gov. http://dx.doi.org/10.1037/e607972011-001

Paul, E. L. (Ed.), (2002). *Taking sides: Clashing views on controversial issues in sex and gender* (2nd ed.). Guilford, CT: McGraw-Hill/Dushkin.

Pearson, K. (1911). On a correction needful in the case of the correlation ratio. *Biometrika, 8,* 254–256. http://dx.doi.org/10.2307/2331454

Pierce, C. A., Block, R. A., & Aguinis, H. (2004). Cautionary note on reporting eta-squared values from multifactor ANOVA designs. *Educational and Psychological Measurement, 64,* 916–924. http://dx.doi.org/10.1177/0013164404264848

Plass, J. L., Moreno, R., & Brünken, R. (Eds.). (2010). *Cognitive load theory*. New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511844744

Pyc, M. A., Agarwal, P. K., & Roediger, H. L. (2014). Test-enhanced learning. In V. A. Benassi, C. E. Overson, & C. M. Hakala (Eds.), *Applying science of learning in education: Infusing psychological science into the curriculum*. Washington, DC: APA Society for the Teaching of Psychology.

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

Pyc, M. A., & Rawson, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. *Science, 330,* 335. http://dx.doi.org/10.1126/science.1191465

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Ravitch, D. (2009, September 15). Critical thinking? You need knowledge. *The Boston Globe*. Retrieved from http://archive.boston.com/bostonglobe/editorial_opinion/oped/articles/2009/09/15/critical_thinking_you_need_knowledge/

Rawson, K. A. (2015). The status of the testing effect for complex materials: Still a winner. *Educational Psychology Review, 27,* 327–331. http://dx.doi.org/10.1007/s10648-015-9308-4

Rice, W. R. (1989). Analyzing tables of statistical tests. *Evolution; International Journal of Organic Evolution, 43,* 223–225. http://dx.doi.org/10.1111/j.1558-5646.1989.tb04220.x

Roediger, H. L., Agarwal, P. K., Kang, S. H. K., & Marsh, E. J. (2010). Benefits of testing memory: Best practices and boundary conditions. In G. M. Davies & D. B. Wright (Eds.), *New frontiers in applied memory* (pp. 13–49). Brighton, UK: Psychology Press.

Roediger, H. L., 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,* 382–395. http://dx.doi.org/10.1037/a0026252

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

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

Rohrer, D., & Pashler, H. (2010). Recent research on human learning challenges conventional instructional strategies. *Educational Researcher, 39,* 406–412. http://dx.doi.org/10.3102/0013189X10374770

Rohrer, D., & Taylor, K. (2006). The effects of overlearning and distributed practise on the retention of mathematics knowledge. *Applied Cognitive Psychology, 20,* 1209–1224. http://dx.doi.org/10.1002/acp.1266

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

Rourke, J. T. (Ed.), (1987). *Taking sides: Clashing views on controversial issues in world politics* (1st ed.). Guilford, CT: Dushkin Publishing Group.

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

Schneider, W., Eschman, A., & Zuccolotto, A. (2007). *E-prime 2 user's guide*. Pittsburgh, PA: Psychology Software Tools.

Skinner, C. H., Fletcher, P. A., Wildmon, M., & Belfiore, P. J. (1996). Improving assignment preference through interspersing additional problems: Brief versus easy problems. *Journal of Behavioral Education, 6,* 427–436. http://dx.doi.org/10.1007/BF02110515

Sweller, J. (2010). Cognitive load theory: Recent theoretical advances. In J. Plass, R. Moreno, & R. Brünken (Eds.), *Cognitive load theory* (pp. 29–47). New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511844744.004

Taylor, K., & Rohrer, D. (2010). The effects of interleaved practice. *Applied Cognitive Psychology, 24,* 837–848. http://dx.doi.org/10.1002/acp.1598

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

Ward, D. (2007). eInstruction: Classroom performance system [Computer software]. Denton, Texas: EInstruction Corporation.

Wheeler, M. A., Ewers, M., & Buonanno, J. F. (2003). Different rates of forgetting following study versus test trials. *Memory, 11,* 571–580. http://dx.doi.org/10.1080/09658210244000414

Wildmon, M. E., Skinner, C. H., & McDade, A. (1998). Interspersing additional brief, easy problems to increase assignment preference on mathematics reading problems. *Journal of Behavioral Education, 8,* 337–346. http://dx.doi.org/10.1023/A:1022823314635

Willingham, D. T. (2009). *Why don't students like school: A cognitive scientist answers questions about how the mind works and what it means for the classroom*. San Francisco, CA: Jossey-Bass.

Appendix A

Counterbalancing Orders

Experiment 1

	Welfare	Vaccines	Multicul	Biotech	SexDiff	Lincoln	Superfund	WWII
1, 2	Study Twice	Study Twice	Fact	Fact	Study Once	Study Once	Higher Order	Higher Order
3, 4	Study Once	Study Once	Higher Order	Higher Order	Fact	Fact	Study Twice	Study Twice
5, 6	Higher Order	Higher Order	Study Once	Study Once	Study Twice	Study Twice	Fact	Fact
7, 8	Fact	Fact	Study Twice	Study Twice	Higher Order	Higher Order	Study Once	Study Once

Note. Odd counterbalancing orders received final fact tests first, alternating with final higher order tests. Even orders received final higher order tests first, alternating with final fact tests.

Experiment 2

	Welfare	Vaccines	Multicul	Biotech	SexDiff	Lincoln	Superfund	WWII
1, 2	Fact 2X	Fact 2X	Mixed (H-F)	Mixed (H-F)	Higher 1X	Higher 1X	Higher 2X	Higher 2X
3, 4	Mixed (F-H)	Mixed (F-H)	Fact 2X	Fact 2X	Higher 2X	Higher 2X	Higher 1X	Higher 1X
5, 6	Higher 1X	Higher 1X	Higher 2X	Higher 2X	Mixed (H-F)	Mixed (H-F)	Fact 2X	Fact 2X
7, 8	Higher 2X	Higher 2X	Higher 1X	Higher 1X	Fact 2X	Fact 2X	Mixed (F-H)	Mixed (F-H)

Note. Odd counterbalancing orders received final fact tests first, alternating with final higher order tests. Even orders received final higher order tests first, alternating with final fact tests.

Experiment 3

	Russian Revolution	World War II
Set A (three class sections)	Higher Order Only	Fact + Higher Order Mix
Set B (three class sections)	Fact + Higher Order Mix	Higher Order Only

(*Appendices continue*)

Appendix B

Sample Questions

Experiments 1 and 2

Session	Fact question	Higher order question
Retrieval practice	Which is the primary reason the "yes" author is against welfare programs?	Which statement is an accurate evaluation or summary of the "yes" author's views?
	1. *Welfare programs don't benefit recipients or taxpayers*	1. Welfare programs can never work, because they are always too expensive
	2. Welfare programs create dependence for recipients	2. Welfare programs are harmful, because they make bad situations even worse
	3. Welfare programs are too expensive for taxpayers	3. *Welfare programs could work, but they rarely meet the needs of the people*
	4. Welfare programs are not the government's responsibility	4. Welfare programs waste taxpayer money on people who don't really need help
Final test	The "yes" author is against welfare programs, largely because	The "yes" author would agree with which statement?
	1. Welfare programs are too expensive for taxpayers	1. *Welfare programs could work, but they rarely meet the needs of the people*
	2. *Welfare programs don't benefit recipients or taxpayers*	2. Welfare programs waste taxpayer money on people who don't really need help
	3. Welfare programs are not the government's responsibility	3. Welfare programs can never work, because they are always too expensive
	4. Welfare programs create dependence for recipients	4. Welfare programs are harmful, because they make bad situations even worse
Retrieval practice	The "no" author argues that vaccines may always carry some amount of risk, but that this risk	Which author would agree with the following statement? "The ends justify the means."
	1. *Is a possibility with any medical procedure*	1. The "yes" author
	2. Is too small to be of concern to the community	2. *The "no" author*
	3. Should be of concern to scientists, not parents	3. Both authors
	4. Is less than the likelihood of a disease epidemic	4. Neither author
Final test	The "no" author argues that vaccine risk	Which author would agree with the following statement? "The achieved outcome is more important than the process along the way."
	1. Should be of concern to scientists, not parents	1. *The "no" author*
	2. *Is a possibility with any medical procedure*	2. Neither author
	3. Is less than the likelihood of a disease epidemic	3. Both authors
	4. Is too small to be of concern to the community	4. The "yes" author

(*Appendices continue*)

Experiment 3

Session	Fact question	Higher order question
Retrieval practice and final test	Why were Nicholas II and the Duma in constant conflict? A. *Because the Duma wanted to help the poor* B. Because Nicholas II wanted to help the poor C. Because the Duma wanted to support communism D. Because Nicholas II wanted control of all of Russia's power	Based on what you know about Nicholas II, how would he treat poor people? A. He would share some power with the poor B. He would help the poor C. He would take money away from the poor D. *He would ignore the poor*
Retrieval practice and final test	Under Stalin, how would you describe everyday life for the Russian people? A. *Stalin controlled all aspects of people's lives* B. People were free to do whatever they wanted C. Stalin forced all people to go to church D. People were allowed to choose their careers	Which person would agree with the following statement? "People are the most productive when they are told what to do by one person, instead of listening to many people or doing what they want." A. Nicholas II B. Lenin C. *Stalin* D. Alexander II
Retrieval practice and final test	What did Franklin Roosevelt do during World War II? A. He dropped an atomic bomb on Japan B. He killed Adolf Hitler C. He joined the Axis war effort D. *He declared war on Japan*	Based on what you know about Franklin Roosevelt, what would he do if Spain attacked the U.S.? A. He would surrender to Spain B. He would negotiate with Spain C. *He would attack Spain in return* D. He would drop an atomic bomb on Spain
Retrieval practice and final test	Why did Hitler join forces with Japan? A. *So they could work together to expand their empires* B. So they could both take over the United States C. So Germany could build an army base in Japan D. So Japan wouldn't join the Allied Forces	Which statement is an accurate summary of Hitler's views? A. By invading the Soviet Union, Germany can increase food production B. By invading the Soviet Union, Germany can create a master race C. *By invading the Soviet Union, Germany can expand its empire* D. By invading the Soviet Union, Germany can strengthen its military

Note. Multiple-choice alternatives remained identical across initial quizzes and final tests for all experiments. The correct answers for each question are italicized. All materials are included in the online supplementary material.

Received June 14, 2017
Revision received March 7, 2018
Accepted March 7, 2018 ■