Review of Educational Research
June 2017, Vol. 87, No. 3, pp. 659–701
DOI: 10.3102/0034654316689306

Rethinking the Use of Tests: A Meta-Analysis of Practice Testing

Olusola O. Adesope
Washington State University

Dominic A. Trevisan
Simon Fraser University, Canada

Narayankripa Sundararajan
Washington State University

The testing effect is a well-known concept referring to gains in learning and retention that can occur when students take a practice test on studied material before taking a final test on the same material. Research demonstrates that students who take practice tests often outperform students in nontesting learning conditions such as restudying, practice, filler activities, or no presentation of the material. However, evidence-based meta-analysis is needed to develop a comprehensive understanding of the conditions under which practice tests enhance or inhibit learning. This meta-analysis fills this gap by examining the effects of practice tests versus nontesting learning conditions. Results reveal that practice tests are more beneficial for learning than restudying and all other comparison conditions. Mean effect sizes were moderated by the features of practice tests, participant and study characteristics, outcome constructs, and methodological features of the studies. Findings may guide the use of practice tests to advance student learning, and inform students, teachers, researchers, and policymakers. This article concludes with the theoretical and practical implications of the meta-analysis.

Keywords: practice test, testing effect, retrieval practice, meta-analysis, systematic review

Johnny comes home from school exhausted. He's scheduled to take five tests within the next few days (American literature, C++ programming, U.S. and Global Economics, Calculus, and Forensic Science), and results will determine whether he can graduate. Despite spending hours each night preparing for exams, he becomes overwhelmed grappling with complex topics. "Why do we have tests?" Johnny exclaims to his parents. "How do I study for these tests? I don't know!" Johnny's parents notice his frustration and are concerned that he's considering dropping out of high school, since he struggled to make it to Grade 12.

"How can we help our only child?" ask the concerned parents, unknowingly rephrasing the question teachers are asking: "How can we help these kids learn and score better on these tests?"

The scenario above reflects the changing school climate in the United States, in which tests are increasingly used to make high-stakes decisions. Policy shifts such as the No Child Left Behind Act and the Common Core State Standards have added considerable pressure to the use and misuse of tests (Coburn, Hill, & Spillane, 2016; Marsh, Roediger, Bjork, & Bjork, 2007; Rothman, 2011; Vinovskis, 2008). Increasingly summative tests are used to make high-stakes decisions such as school accountability and funding, merit pay for teachers based on student performance on standardized tests, school readiness, students' promotion to higher grades, and admission into colleges. Consequently, testing and the issues surrounding tests (e.g., scoring, bias, reliability, validity, etc.) have attracted large debates and scholarly attention. This trend has led some policymakers and educators to view tests mainly as summative assessment tools to measure students' mastery of skills and knowledge (Marsh et al., 2007). Regrettably, the heavy emphasis on tests as summative assessment tools used to make high-stakes decisions often obscure some very important function of tests. These include the opportunity to use test results for low-stakes formative assessments and provide students with feedback on their strengths and weaknesses, as well as helping teachers improve their instruction. When test results are used in this formative way, teachers and students are provided with additional resources that can help adapt teaching and learning to improve student achievement (Black & Wiliam, 1998). Hence, in this article, we do not advocate for the current wave of high-stakes testing and accountability initiatives. Instead, we seek to rigorously examine whether and how low-stakes practice tests can be used to improve learning.

Several decades before the national spotlight on testing and assessment, a phenomenon called the *testing effect* was coined. *The testing effect* is a cognitive psychology term referring to the finding that taking practice tests on studied material promotes greater subsequent learning and retention on a final test compared to more common study strategies (Roediger & Karpicke, 2006a). The testing effect shows that "retrieval processes used when taking a test have powerful effects on learning and long-term retention" (Agarwal, Karpicke, Kang, Roediger, & McDermott, 2008, p. 861). The testing effect we examine in this meta-analysis is different from a less common use of the term *testing effect* (also referred to as "testing threat") used in experimental psychology, which describes a threat to the internal validity of pre–post research design (Fraenkel, Wallen, & Hyun, 2006). In this article, we use the term *testing effect* exclusively to refer to the assumption that practice tests and other types of retrieval practices yield greater learning benefits than other common study strategies. Throughout this article, we use the terms *practice tests*, *testing effect*, and *retrieval practice* interchangeably to refer to the use of practice tests as a mechanism for enhancing retention and learning (Roediger & Butler, 2013).

Although the convergent finding in the extant literature is that students learn and recall information better after studying the material and taking a practice test compared to only restudying the material, the magnitude of the effect differs across studies and depends on several factors. These factors include and are not

limited to (a) test format (multiple-choice versus short-answer), (b) individual differences among participants, and (c) whether studies were conducted in laboratory environments or occurred in classrooms with the use of materials relevant to the curriculum. Researchers have noted the variability in findings in testing effect studies and remarked on the need to identify the optimal uses of tests for learning (Carpenter & Pashler, 2007; Marsh et al., 2007; Roediger & Karpicke, 2006b). Our main goal in this work is to synthesize existing research using a meta-analysis to explore the different conditions under which practice tests can enhance or inhibit learning.

Specifically, this meta-analysis addresses the following research questions:

Research Question 1: What are the learning effects of taking a practice test compared with other learning conditions?
Research Question 2: Do various features of practice tests foster different learning benefits on a final test?
Research Question 3: To what degree would the testing effect vary based on whether feedback was (or was not) provided during the initial practice test?
Research Question 4: How does the testing effect vary when used for learning in different settings (classroom or laboratory), in educational levels, and with different learning outcome constructs?
Research Question 5: How are effect sizes moderated by contextual and methodological features of the research?

Previous Reviews of Testing Effects

The testing effect has been revisited repeatedly over several decades. Although research on testing has been conducted before the turn of the 19th century, Roediger and Karpicke (2006b) observed that the first large-scale study on testing effects was conducted by Gates in 1917. Since then, the effect has been studied sporadically. Recently, researchers have revitalized interest in this phenomenon, with many studies over the past two decades (e.g., Chan, McDermott, & Roediger, 2006; Glover, 1989; Johnson & Mayer, 2009; McDaniel, Agarwal, Huelser, McDermott, & Roediger, 2011; McDaniel & Fisher, 1991; Roediger & Karpicke, 2006a; Rohrer & Pashler, 2010). However, relatively few reviews and meta-analyses have been conducted to date.

A prior meta-analysis of research on the effects of frequent classroom testing was conducted by Bangert-Drowns, Kulik, and Kulik (1991). They reviewed 35 classroom studies (22 published and 13 unpublished) comparing a frequently tested experimental group of students with a control group that received considerably fewer tests. A majority of the studies focused on social studies (17 studies). There were six studies each in mathematics, science, and other domains. Compared with students who took few or no tests, the use of frequent tests offered gains in learning averaging about $d = 0.23$ standard deviations. Effects of practice tests were found to be higher in mathematics ($d = 0.28$) than in other domains. The meta-analysis also showed that with increased test frequency, student achievement improved at a progressively slower rate with each additional test. More recently, Phelps's (2012) summary of research, consisting of quantitative and qualitative studies, on the testing effect found a moderate to large effect in support

of using practice tests for learning. Phelps also found that the effects of practice tests were amplified when feedback was included with tests.

In a selective narrative review on the power of testing in improving memory and retention, Roediger and Karpicke (2006b) reviewed some historical and contemporary findings on testing effects. They observed that practice tests are beneficial in free recall of information, and that repeated testing of the learning material attenuates forgetfulness and significantly improves long-term retention. They found that testing effects also exist in cued-recall and paired-associate learning, in which participants learn paired items such that presentation of one item of the pair evokes memory recall of the other item. In fact, repeated tests were found to provide greater benefits. The benefits of testing extended beyond laboratory experiments, as findings demonstrated positive effects of testing with educational materials and studies conducted in the classroom (Karpicke & Aue, 2015), although results were mixed on whether essay-type or short-answer tests produce greater benefits on later tests than multiple-choice tests. More recently, Rowland (2014) authored a meta-analysis to understand the theoretical underpinnings of studies that examined the testing effect. Rowland's meta-analysis examined the effects of testing versus restudy on retention to understand how findings align with existing theoretical accounts. Rowland's meta-analysis did not examine other comparison conditions outside of restudy.

The Need for a New Meta-Analysis on the Testing Effect

Due to a lack of comprehensive evidence-based meta-analysis, there is a need for comprehensive understanding of conditions under which the testing effect enhances or inhibits learning. Empirical guidance toward a theory of how learning processes are affected by the testing effect will ultimately benefit policymakers, researchers, teachers, and students. Many policy and educational shifts have occurred, including the more recent Every Student Succeeds Act (Executive Office of the President, 2015), affecting the users and uses of tests for learning. Many studies have been published since the meta-analysis by Bangert-Drowns et al. (1991), and thus a reevaluation of the testing effect research base is needed. Our meta-analysis differs significantly from previous reviews and meta-analyses in the following ways:

- It explores the moderating influences of different practice and final test formats, as well as corrective feedback on learning not examined in Bangert-Drowns et al.'s (1991) meta-analysis;
- It examines both classroom and laboratory studies on the effects of practice tests on learning, whereas Bangert-Drowns et al.'s meta-analysis only investigated classroom-based studies;
- The present meta-analysis is different from Phelps (2012) in at least four ways. First, Phelps (2012) conducted an extensive review of quantitative, survey, and qualitative studies on testing effects published up to 2010 and provided results separately for each, while our study focuses on quantitative studies published up to 2015. Indeed, of the 282 independent effect sizes that we extracted and included in our meta-analysis, 84 were published between 2011 and 2015. Second, Phelps (2012) examined both low-stakes and high-stakes testing effects while our meta-analysis only examined the effects of low-stakes practice tests. Third, our meta-analysis

reported key methodological information such as study selection criteria, index of inter-rater reliability (Kappa) and presented weighted mean effect sizes with concomitant statistics such as standard errors, 95% lower and upper confidence intervals, and tests of heterogeneity (e.g., Q and I^2 statistics). Researchers have recommended that such information be reported as they help readers understand and interpret the precision and heterogeneity of the effect sizes so as to provide more reliable and valid inferences from meta-analyses (Ahn, Ames, & Myers, 2012; Borenstein, Hedges, Higgins, & Rothstein, 2009; Harwell & Maeda, 2008). Fourth, while we acknowledge the extensive review and report of several moderators by Phelps, our meta-analysis examined several moderators not reported by Phelps. These moderators include, but are not limited to, comparison treatment, practice and final test format, transfer appropriate processing, study settings, measure reliability, prior knowledge, treatment fidelity;

- Over the 20 years since Bangert-Drowns et al.'s meta-analysis, major methodological advances have been made in meta-analytical techniques that permit more detailed investigation of variability in findings;
- Many studies on the use of tests have not been summarized after major educational policy changes, such as implementation of the No Child Left Behind Act of 2001 and the Common Core State Standards in 2010;
- Unlike Roediger and Karpicke's (2006b) narrative review that examined a representative number of studies, this meta-analysis uses a systematic, comprehensive approach to recover, synthesize, and analyze studies on testing effects and report the effect of testing under varying conditions;
- We included classroom studies and more comparison conditions than Rowland (2014), providing a more comprehensive understanding of the testing effect phenomenon.

This work expands knowledge in this area with a rigorous meta-analytical investigation of the testing effect, as well as an examination of theoretically and empirically grounded moderating variables to identify how the testing effect differs across settings and study designs.

Factors That Influence Learning Through Practice Tests

This section discusses various factors that may potentially moderate learning with practice tests leading to different results. We acknowledge that the factors discussed below are not exhaustive. However, we thoughtfully selected these factors based on the degree to which they are emphasized or experimentally manipulated in the literature.

Nature of Comparison Group

Most testing effect studies compare the benefits of practice tests to a comparable *nontesting* learning activity (e.g., rereading, concept mapping, etc.). Rereading, or reexposure to initial learning material without direct instruction on how to study it, is most commonly used, since it replicates what students typically do during self-directed study sessions (Kornell & Bjork, 2007; Kornell & Son, 2009). Other testing effect studies compare practice tests to a no-exposure comparison condition or a filler activity that is unrelated to the initial learning material.

The ways that comparison groups affect the strength of testing effects have important methodological implications for future research that explores the benefits of practice tests. To make a convincing case that retrieval is an effective learning strategy beyond an additional practice opportunity, it is important to compare practice tests to other nontesting study activities matched for time and content. In this meta-analysis, we categorize comparison conditions into two basic groups: (a) any condition that receives an additional representation to the learning material but does not receive practice tests (e.g., additional study, rereading, practice, concept mapping) is categorized into a restudy group and (b) any condition that does not receive an additional exposure to the material (e.g., no intervention or an unrelated filler activity) is categorized into a no-activity group. This categorization may mimic students' study habits for practical relevance to learners who want to improve their efficiency.

Test Format and Transfer-Appropriate Processing

Studies on testing effects use a variety of different test formats, including free-recall, cued-recall, multiple-choice, short-answer, and recognition. Free-recall tests usually present a word list in the initial learning phase of the experiment, and ask participants in the testing condition to list all of the words that they can remember from the initial presentation. On the other hand, cued-recall tests usually consist of a presentation of a word pair in the presentation condition (e.g., flower–daisy) and with only the presentation of a cue in the retrieval practice condition (e.g., flower–_____). Similarly, in a multiple-choice test, students are first presented with a cue (question), and then with several choices for the target (answer). The only difference between multiple-choice and cued-recall tests is that in cued-recall tests, no choices for the target are offered after the cue is presented. This requires more cognitive effort than a multiple-choice test (Duchastel & Nungester, 1981). Short-answer tests are another form of a cued-recall test, but they differ based on the type of cue and target used. They use a similar format to cued-recall but use questions and answers instead (e.g., What is the capital of Latvia? _____). Last, in a typical recognition test, a word list is presented in an initial learning session. Then, in the testing condition, participants are presented with a new word list and simply have to recall whether a word was presented previously (see Carpenter & DeLosh, 2006; Jacoby, Wahlheim, & Coane, 2010; Karpicke & Zaromb, 2010, Exp. 3).

There are at least two important issues to consider regarding test format in testing effect research. First, is it important for the test format of the practice test learning condition and the final test to be identical? This question relates to the concept of *transfer-appropriate processing* (TAP), the principle that performance on any given task will be highest if the characteristics of the learning procedure are similar to the characteristics of the assessment procedure (see Bransford, Franks, Morris, & Stein, 1979). In fact, the reason that a practice test is such a powerful test preparation tool may be because it imitates the mental processes to be performed on the actual examination. Extending this logic, it is worth investigating whether testing effects are maximized if practice and test formats are identical.

A second issue worthy of consideration is whether some practice test formats are optimal for subsequent retention, regardless of the format of the final assessment test. Some studies have attempted to explore how the strength of testing effects are

influenced by how cognitively demanding the retrieval processes are during the practice test phase of an experiment. Since multiple-choice and short-answer tests are most commonly used in educational settings and the different levels of processing required between the two tests are obvious (having choices or not having choices, respectively), studies comparing the effects of different levels of processing often compare short-answer and multiple-choice tests (e.g., Kang, McDermott, & Roediger, 2007). For instance, Kang et al. (2007) revealed that students who took a short-answer practice test outperformed students who took a multiple-choice practice test on the final test, regardless of whether the final test was short-answer or multiple-choice. This result offers two important considerations. First, consistent with previous findings (e.g., R. A. Bjork & Whitten, 1974; Butler & Roediger, 2007; Glover, 1989; McDaniel, Anderson, Derbish, & Morrisette, 2007), the benefits of retrieval practice are more salient, as levels of processing during retrieval become more demanding. Second, these findings suggest that the cognitive demand of the practice test format may be more relevant, particularly for retention, than the principle of TAP (see also Carpenter & DeLosh, 2006). Our meta-analysis examines whether TAP is a significant moderator of the testing effects.

Feedback

Experiments on testing effects often manipulate a variable known as corrective feedback, which involves a phase of the experiment in which a participant's answers to initial practice tests are indicated as correct or incorrect. Other studies, rather than providing corrective feedback, merely reexpose the initial learning material to participants after the practice test so that they can restudy the material and perhaps think back to the practice test to see which items were answered correctly. Providing feedback in practice tests may act as formative assessment of student learning by confirming correct answers and providing information on questions that were answered incorrectly, thereby increasing the positive effects of practice tests. When corrective feedback is given on practice tests, students are afforded an opportunity to correct errors and retain correct answers on a later test (Butler & Roediger, 2008). Although some researchers have found that students who receive corrective feedback on exams outperform those who do not receive feedback (e.g., Butler & Roediger, 2008; McDaniel & Fisher, 1991), some researchers did not find significant differences between conditions that received or did not receive corrective feedback. Butler and Roediger (2007) found a considerably high level of correct answers on the practice tests, which may explain their finding that feedback did not help final retention.

Some studies show that feedback can even be redundant in certain situations. For example, feedback on high-confidence correct answers can be an inefficient use of time and cognitive energy (Hays, Kornell, & Bjork, 2010; Karpicke & Roediger, 2008). Given the large amount of information students are expected to learn in various educational settings, it is important they use their study time as efficiently as possible. However, other studies found that feedback improves retention even for items answered correctly on the practice tests (Butler, Karpicke, & Roediger, 2008), especially when questions were answered correctly with low confidence. Given the inconsistent findings on the benefits of feedback, it is important to investigate related variables that may facilitate or inhibit its benefits on test performance (Hattie & Timperley, 2007). For example, some studies have examined the different effects

of immediate versus delayed feedback. A meta-analysis by Kulik and Kulik (1988) concluded that immediate feedback tends to be more beneficial than delayed feedback in applied classroom settings, although the opposite was found for controlled laboratory studies. Butler, Karpicke, and Roediger (2007) suggested that this may be because students in a classroom will pay more attention to immediate feedback than delayed feedback. Delayed feedback may be in the form of a corrected quiz the day after taking the quiz; there is no guarantee that the students will go over each question rather than simply viewing their percentage correct on top of the page. Conversely, in a laboratory setting, a delayed feedback condition would be carefully controlled, and participants would be expected to study that feedback for a set amount of time. Consistent with Kulik and Kulik's (1988) findings, Butler et al.'s (2007) laboratory study observed better long-term retention for their delayed feedback conditions in two experiments. Our meta-analysis aims to examine the potential moderating effects of feedback.

Study Setting

Testing effect studies are typically conducted either in classrooms or in controlled laboratory settings. A goal of this meta-analysis is to examine the degree to which study setting (i.e., classroom vs. laboratory) moderates the testing effect. This is important, as it provides a rationale for whether educators should incorporate practice tests into their instructional strategies. Although the vast majority of studies analyzed in this meta-analysis are laboratory-based, about 11% of the coded experiments took place in a classroom setting. We classified a study as a classroom study only if the materials were part of the curriculum. Although the positive learning effects of testing in laboratory settings have been well documented, few studies have attempted to analyze these effects in a real classroom (see Cranney, Ahn, McKinnon, Morris, & Watts, 2009; Glover, 1989; Mayer et al., 2009; McDaniel et al., 2011; McDaniel, Thomas, Agarwal, McDermott, & Roediger, 2013; McDaniel, Wildman, & Anderson, 2012).

Several key differences exist between classroom and laboratory experiments. Students in classrooms have higher levels of extrinsic motivation such as grades and class promotion (Agarwal et al., 2008; Phelps, 2012), as well as intrinsic motivation, the belief that what they are learning is important (Delaney, Verkoeijen, & Spirgel, 2010). When examining testing effects in a controlled laboratory setting, one disadvantage concerns the levels of incentives or motivation participants have to learn the material and perform well on final retention tests. Phelps (2012) found a positive correlation between stakes of test performance and achievement. In a classroom setting, the motivation is, of course, receiving high marks on an exam to receive a good grade in the class. In laboratory studies, often an incentive is included, such as a small monetary payment or extra credit for undergraduate courses. However, these types of incentives typically do not emulate the stakes of a classroom test, as money or extra credit is awarded regardless of test performance or levels of effort. Another difference is that students in a classroom are expected to master more material for subsequent tests than students recruited into a laboratory study. Furthermore, the amount of time participants actually study is typically tightly controlled in a laboratory study, while study preparation varies considerably in classroom-based studies (Roediger & Karpicke, 2006b).

Education Level and Sample Characteristics

Testing effect studies have been conducted with participants of diverse age groups in diverse populations that range from elementary or preschool-aged students (e.g., Bouwmeester & Verkoeijen, 2011; Fritz, Morris, Nolan, & Singleton 2007) to elderly adults (e.g., Bishara & Jacoby, 2008; Tse, Balota, & Roediger, 2010). Most testing effect studies have been conducted in unimpaired populations for all age and education levels, but some have examined atypical populations associated with memory dysfunction (e.g., Sumowski, Chiaravalloti, & DeLuca, 2010). Roediger and Karpicke's (2006b) review suggested that frequent low-stakes classroom testing might elevate educational achievement at all levels of education. Nevertheless, results from our meta-analysis identify which population characteristics and education-level testing effects yield the strongest learning benefits in order to provide recommendations for educators on the use of retrieval practice in the classroom.

Experimental Design

Two common experimental designs used in research on testing effects include within-subjects and between-subjects designs. Although many studies use mixed factorial designs (in which some variables are manipulated within-subjects and others are manipulated between-subjects), we classified studies as within-subjects or between-subjects based only on how the learning condition was manipulated. However, when we collapsed multiple effect sizes from the same sample in order to maintain statistical independence, we referred to them as a "mixed" category. With a between-subjects design, separate groups of participants will form the practice test and control conditions. One drawback of this design is that it typically requires twice as many participants as a within-subjects design to achieve the same level of power. There is also the risk that differentiating retention scores could have been caused by prior knowledge differences rather than the nature of the learning condition. Of course, this can be minimized if the research controls for prior knowledge differences using randomization, participant matching, or another method.

In a within-subjects design, each participant takes part in the experimental (practice tests) and control (nontesting) conditions, by using retrieval practice for some items and a different learning strategy for other items. The methodological advantage of the within-subjects design is that the design eliminates the potential confound of prior group differences. Thus, any observed differences on retention scores are presumed to be explained by manipulation of the learning condition. A possible methodological disadvantage concerns the potential that retrieval practice on some items will affect the memorization of other nontested material (Chan, 2009, 2010; Chan et al., 2006; Spitzer & Bäuml, 2009) in processes known as retrieval-induced forgetting (RIFO) and retrieval-induced facilitation (RIFA). RIFO occurs when the act of retrieval triggers mental interference of other unrelated information, which in turn becomes weeded out and forgotten in order to recall the target information (see M. C. Anderson, Bjork, & Bjork, 1994; Murayama, Miyatsu, Buchli, & Storm, 2014). Consistent with theories of associated memory proposed by J. R. Anderson, Reder, and Lebiere (1996) and Raaijmakers and Shiffrin (1981), RIFA occurs when recall of certain information activates related information. For example, recalling the day that Richard Nixon announced his resignation may activate recall of details about the Watergate scandal leading to his impeachment. In sum, testing effects observed from within-subjects designs could

be confounded by RIFO or RIFA. Considering the differences between within-subjects and between-subjects experimental designs, one goal of this study will be to examine how different experimental designs moderate the strength of testing effects. The next section describes our methods, including how studies were searched, selected, and analyzed.

Method

Our approach to conducting a meta-analysis is consistent with well-established review protocols, applying the following procedures to the collection and synthesis of research (Adesope, Lavin, Thompson, & Ungerleider, 2010; Adesope & Nesbit, 2012; Cooper, Hedges, & Valentine, 2009; Lipsey & Wilson, 2001; Nesbit & Adesope, 2006).

Study Selection Criteria and Search Strategies

Following a preliminary examination, of empirical studies and reviews of literature, we developed criteria to capture all relevant studies investigating the testing effect. Studies were deemed eligible to be included in the meta-analysis if they met all the following criteria: (a) contrast the effects of taking a practice test with the effects of restudying or other learning strategies; (b) report measurable cognitive outcomes such as recall or transfer; (c) report sufficient data to allow for effect size extractions; (d) publicly available through databases, journals, or library archives; and (e) random assignment, within-subjects designs, or other means to control for preexisting group differences such as pretests or other matching procedures in a between-subjects design.

Studies were located in 2014 through comprehensive and systematic searches involving several databases and search strategies. We used the query: *testing effect** OR *test effect** OR *retrieval practice** to conduct a comprehensive and systematic search on the following electronic databases: ERIC, PsycARTICLES, PsycINFO, and Web of Science. The reference sections of a number of articles that investigated the testing effect were also searched to recover studies not captured with the database searches (e.g., Bangert-Drowns et al., 1991; Glover, 1989; McDaniel et al., 2011; Roediger & Karpicke, 2006b; Rohrer & Pashler, 2010). The search procedure returned a total of 1,717 studies.

Document Retrieval, Secondary Review, and Data Extraction and Analysis

We adopted two selection phases to determine whether articles returned by the searches should be included or excluded from the meta-analysis. In the first phase, both the first and second authors read the titles, abstracts, and keywords of these 1,717 studies for possible inclusion by applying the selection criteria. Studies identified as duplicates and those not meeting the selection criteria were excluded. For borderline cases in which the abstract did not provide sufficient information to include or exclude the study based on our selection criteria, the first and second authors made a joint decision. In several cases, we read the method, procedure, and data collection sections of such borderline studies to obtain more information that helped us retain or exclude the paper. In the first phase, 334 studies met all inclusion criteria. In the second phase, we developed a coding form and read full-text copies of all 334 articles that met inclusion criteria in the first phase to determine eligibility based on specified inclusion

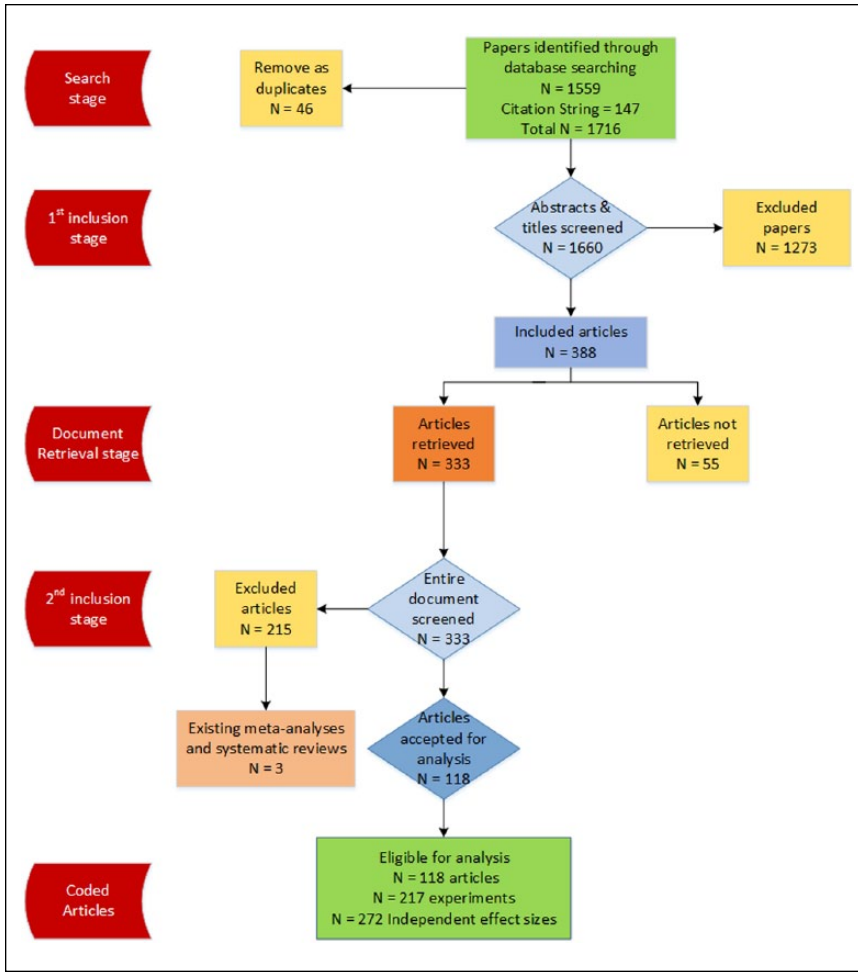

FIGURE 1. *Flowchart for selection of studies.*

criteria. Researchers worked together to ensure a rigorous coding process. At the start of the coding phase, researchers independently coded several similar studies and met to verify their coding. After they had established reliability of the coding process, data from the articles that met the inclusion criteria at the second phase were extracted. When variables were not reported in the article, they were coded as "not reported" and associated statistics displayed in the tables of results. A reliability check on the extracted data was conducted by the one of the authors, who randomly selected and coded about 30% of all studies included in this meta-analysis. Interrater agreement was high (k = .96). Figure 1 shows the flowchart of how studies were filtered throughout the process of searching for studies to be included in this meta-analysis.

For each outcome measure, we extracted Cohen's *d* effect size. Cohen's *d* effect size is a standardized estimate of the difference in mean scores between students who took a practice test before a final test compared with those who did not take a practice test before taking a final test, divided by the pooled standard deviation of the two groups. Since differential sample sizes across studies may bias the effect size obtained by Cohen's *d*, we used Hedges's *g* to adjust all effect sizes to provide unbiased estimates of effect size (Hedges, 1981; Hedges & Olkin, 1985) using Equation (1) below. When other statistics such as *F* or *t* were provided, we used them to derive effect sizes or to verify the obtained Cohen's *d* (Cooper et al., 2009). We analyzed data using Comprehensive Meta-Analysis 2.2.048 (Borenstein, Hedges, Higgins, & Rothstein, 2008) and SPSS Version 23 for Windows.

$$g = (1 - \frac{3}{4N - 9})d, \quad (1)$$

where *N* is the total number of participants in the experimental (testing) and comparison groups and *d* is the biased Cohen's *d* effect size.

Of the 272 independent effect sizes that were coded for the meta-analysis, 5 effect sizes produced extreme standardized scores ($-3.3 \geq Z \geq 3.3$; $p < .001$) and were thereby identified as outliers. Further examination of those five outlying studies did not reveal any methodological flaws, and hence a decision was made to adjust each effect size toward the nearest other effect size in the distribution, as recommended by Tabachnick and Fidell (2013). We then used the five adjusted effect sizes in all subsequent analyses.

Results

A total of 118 articles yielding 272 independent effect sizes involving 15,427 participants were analyzed on different features and outcomes. Figure 2 shows the distribution of effect sizes. To offer coherence of presentation, the results of the overall and moderator analyses were organized and presented around the research questions. In all reported analyses, a positive weighted mean effect size (*M*) indicated that students benefited from taking practice tests before taking a final test. Tables of results include the number of participants (*N*) in each category, the number of findings (*k*), the weighted mean effect size (*g*) and its standard error (*SE*), the 95% lower and upper confidence intervals, the results of a test of homogeneity (*Q*) with its concomitant degrees of freedom (*df*), as well as the percentage of variability that could be attributed to true heterogeneity or between-studies variability (I^2).

What Are the Learning Effects of Taking a Practice Test Compared With Other Learning Conditions?

Table 1 presents an overall analysis of the weighted mean of all statistically independent effect sizes, as well as the results of learning with practice tests compared with other learning conditions. Under a fixed-effects model, Table 1 shows that the overall weighted mean effect size was moderately large and statistically significant, indicating the effectiveness of learning with practice tests ($g = 0.61$, $p < .001$). However, the overall sample was heterogeneous, $Q(271) = 1405.00$, $p < .001$, $I^2 = .81$. The total variability that could be attributed to true heterogeneity or between-studies variability was 81%, indicating that 81% of the variance

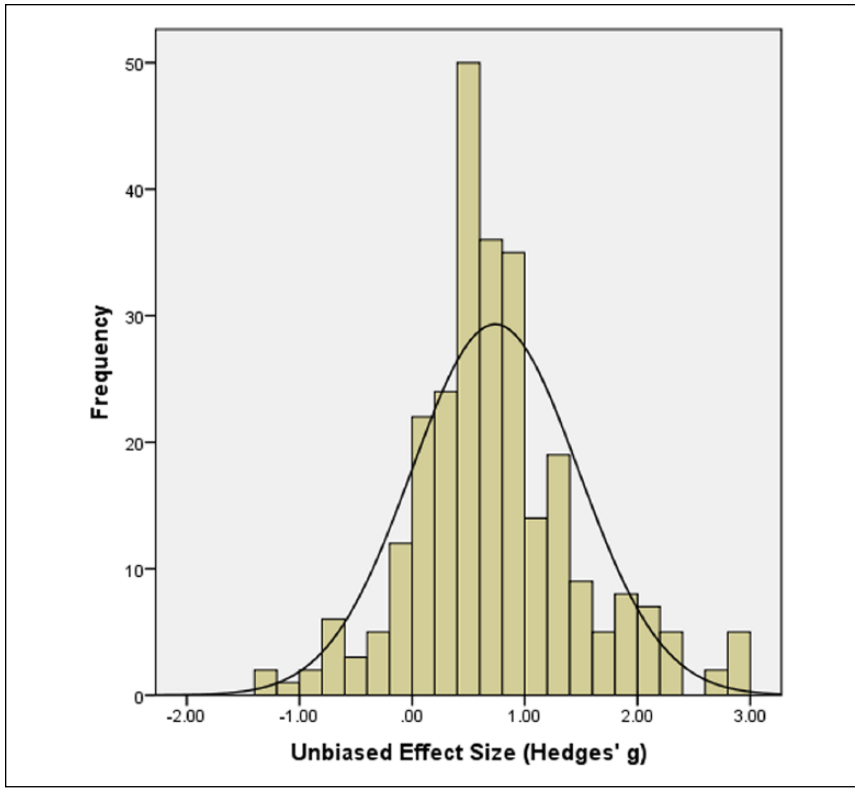

FIGURE 2. *Distribution of 272 independent effect sizes obtained from 118 articles (*M = *0.74,* SD = *0.74).*

could be explained by study-level covariates and 19% of the variance was within-study variance based on sampling error. Generally, this implies a significant variability in the individual effect sizes that constitute the overall result. Hence, moderator analyses were conducted to examine the study features that may have been responsible for the variability. The overall result with a random-effects model showed a larger effect size that was statistically significant ($g = 0.70$, $p < .001$). Many studies in our meta-analysis were conducted by few researchers and mostly used similar operationalizations of the design conditions and approaches to examine the effects of practice tests. We determined that variations in effect sizes were attributed to within-study estimation error—that is, sampling variance (Borenstein, Hedges, Higgins, & Rothstein, 2009; Hedges & Vevea, 1998). Hence, we decided to use the fixed-effects model for all analyses.

Table 1 also lists the breakdown of the comparison treatment, showing that most studies compared the use of practice tests with restudying ($k = 195$). About 30% of the studies ($k = 58$) compared the use of practice tests with filler activities or, in some cases, with no reading activities. Results demonstrate that the use of practice tests was associated with a moderate, statistically significant weighted

TABLE 1

Overall weighted mean effect size

			Effect size		95% confidence interval		Test of heterogeneity			
Grouping	*N*	*k*	*g*+	*SE*	Lower	Upper	*Q*	Degrees of freedom	*p*	I^2 (%)
All (fixed-effects model)	15,427	272	0.61*	0.02	0.58	0.65	1405.00	271	<.001	80.71
All (random-effects model)	15,427	272	0.70*	0.04	0.63	0.78				
Comparison treatments										
Restudying/rereading	10,491	195	0.51*	0.02	0.47	0.55	896.88	194	<.001	78.37
Filler or no activity	3,452	58	0.93*	0.04	0.86	1.01	371.55	57	<.001	84.66
Mixture of restudying and filler activities	1,340	18	0.71*	0.06	0.59	0.82	32.82	17	.01	48.19
Not reported	144	1	0.68*	0.17	0.35	1.01				
Between-levels (Q_B)							103.75	3	<.001	

*$p < .05$.

mean effect size compared to restudying ($g = 0.51$), and a much larger weighted mean effect size ($g = 0.93$) when compared with filler or no activities.

Do Various Features of Practice Tests Foster Different Learning Benefits on a Final Test?

To address this question, we analyzed moderator variables of practice test formats, final test formats, and the number of practice tests. We also examined whether the studies adopted TAP. Table 2 displays weighted mean effect sizes for each the aforementioned moderator variables. Researchers mainly used recall practice tests, either as cued-recall ($k = 134$) or as free-recall tests ($k = 48$). All of the test formats produced moderate to large effect sizes. Specifically, mixed test formats and multiple-choice practice tests produced large effect sizes ($g = 0.80$ and $g = 0.70$, respectively). The between-levels difference was statistically significant, $Q_B(7) = 27.02$, $p < .001$. Post hoc analysis revealed that practice tests with mixed-format tests were associated with higher weighted mean effect sizes than free-recall, cued-recall and short-answer tests, and differed significantly from all three. Post hoc analysis also revealed that multiple-choice practice tests were associated with higher weighted mean effect size than short-answer tests, and differed significantly from them.

Furthermore, findings show that TAP moderated the strength of testing effects. This indicates that testing effects were stronger when the practice and final test formats were identical ($g = 0.63$) compared to when the practice and final test formats were dissimilar ($g = 0.53$). The between-levels variance was statistically significant, $Q_B(3) = 15.73$, $p < .001$. Post hoc analysis revealed that when practice tests were identical to the final test formats, the testing effect had significantly higher weighted mean effect sizes and differed significantly from when practice tests were dissimilar in format to final tests. Results also show that when formats were mixed ($k = 28$), that is, when the practice or final test had both a common format and a different format, the weighted mean effect size was large $g = 0.75$.

Finally, Table 2 shows the weighted mean effect sizes of the effects of practice tests on various forms of final tests. Similar to formats of practice tests, the majority of studies evaluated used cued-recall ($k = 106$) and free-recall ($k = 49$) as final tests. However, unlike practice test formats, final test formats produced small to large effect sizes with more robust testing effects when the final test formats were free-recall, cued-recall, multiple-choice, or short-answer. Due to variability in the testing effect on final test formats, we examined whether the number of practice tests influences the magnitude of the testing effect. Table 2 shows the degree to which the number of practice tests influences learning. One hundred and sixty-five studies required participants to use practice tests only once, while 100 studies required participants to use practice tests at least twice. Results demonstrate that the effects of practice tests were larger when they were used only once ($g = 0.70$) than when they were used twice or more ($g = 0.51$).

To What Degree Would the Testing Effect Vary When the Initial Practice Test Was Given With or Without Feedback?

Table 2 shows that testing effects were robust, whether or not feedback on practice tests was given. The weighted mean effect sizes do not differ significantly when feedback was given ($g = 0.63$; $k = 119$) or not ($g = 0.60$; $k = 153$).

TABLE 2

Weighted mean effect sizes for features of practice test

			Effect size		95% confidence interval		Test of heterogeneity		
Moderator	*N*	*k*	*g*+	*SE*	Lower	Upper	*Q*	Degrees of freedom	I^2 (%)
Format of practice test									
Free-recall	2,592	48	0.62*	0.04	0.54	0.70	262.12*	47	82.07
Cued-recall	6,800	134	0.58*	0.03	0.53	0.63	733.35*	133	81.86
Multiple-choice	1,725	22	0.70*	0.05	0.60	0.80	158.58*	21	86.76
Short-answer	1,628	24	0.48*	0.05	0.38	0.58	97.03*	23	76.30
Mixed-format	1,626	30	0.80*	0.05	0.70	0.91	83.55*	29	65.29
Other	657	10	0.57*	0.08	0.41	0.73	37.15*	9	75.78
Not reported	399	4	0.78*	0.11	0.57	0.99	6.20	3	51.60
Between-levels (Q_B)							27.02*	6	
Format of final test									
Free-recall	2,868	49	0.71*	0.04	0.63	0.79	451.19*	48	89.36
Cued-recall	5,117	106	0.62*	0.03	0.56	0.68	415.25*	105	74.71
Multiple-choice	2,006	25	0.56*	0.05	0.46	0.65	152.70*	24	84.28
Short-answer	1,780	30	0.67*	0.05	0.57	0.76	121.75*	29	76.18
Mixed-format	1,255	25	0.78*	0.06	0.67	0.90	61.24*	24	60.81
Other	1,645	33	0.40*	0.05	0.30	0.50	144.86*	32	77.91
Not reported	756	4	0.48*	0.08	0.33	0.63	21.94*	3	86.33

(continued)

TABLE 2 (continued)

Moderator	*N*	*k*	Effect size *g*+	*SE*	95% confidence interval Lower	Upper	Test of heterogeneity *Q*	Degrees of freedom	I^2 (%)
Between-levels (Q_B)							36.06*	6	
No. of practice tests									
One	9,006	165	0.70*	0.02	0.66	0.74	950.08*	164	82.74
Two or more	5,794	100	0.51*	0.03	0.46	0.56	411.75*	99	75.96
Self-paced	407	4	0.31*	0.10	0.12	0.51	0.66	3	0.00
Mixed (one + more than one)	119	2	0.81*	0.19	0.44	1.18	0.26	1	0.00
Not reported	41	1	0.08	0.31	−0.52	0.68	0.00	0	0.00
Between-levels (Q_B)							42.25*	4	
Transfer-appropriate processing									
Practice and final test formats were identical	9,390	175	0.63*	0.02	0.59	0.67	757.03*	174	77.02
Practice and final test formats were not identical	4,449	69	0.53*	0.03	0.47	0.59	559.24*	68	87.84
Mixed	1,588	28	0.75*	0.05	0.65	0.86	73.56*	27	63.29
Between-levels (Q_B)							15.73*	3	
Feedback on practice test									
No	8,717	153	0.60*	0.02	0.56	0.65	942.48*	152	83.87
Yes	6,710	119	0.63*	0.03	0.58	0.68	462.06*	118	74.46
Between-levels (Q_B)							0.46		

*$p < .05$.

Although there was no significant difference between conditions when feedback was given and when it was not, the weighted mean effect size for feedback was slightly more than when there is no feedback.

How Do Testing Effects Vary When Used for Learning in Different Settings, Designs, and Educational Levels and With Different Outcome Constructs?

Table 3 shows results of moderator analyses based on study settings, experimental designs, educational levels, and learning outcome constructs. We examined the effects of practice tests in classroom settings (k = 30), where learning activities were a part of classroom performance assessment as well as in laboratory settings (k = 223), where learning activities did not contribute to performance assessment. Results show that testing effects do not vary with study settings. Studies conducted in the classroom had a moderately large weighted mean effect size (g = 0.67), similar to those in the laboratory (g = 0.62). Results also show that practice tests produced moderate to large and statistically significant weighted mean effect sizes across different experimental designs.

Table 3 shows that even though the majority of studies were conducted with postsecondary students, studies conducted with primary, secondary, and postsecondary school students all produced moderate to large weighted mean effect sizes (g = 0.64, g = 0.83, and g = 0.60, respectively). The between-levels difference was statistically significant, $Q_B(4) = 50.50$, $p < .001$. Post hoc analysis revealed that the use of practice tests by secondary school students was associated with higher weighted mean effect size and differed significantly from results on postsecondary students' use of practice tests.

The learning outcome variable was coded into four categories: retention, transfer, mixed (retention and transfer), and not reported, a category for studies that did not report their outcome variable. Although most studies used retention tests (k = 243), results show that practice tests were effective, regardless of outcome variables. In other words, the use of practice tests was associated with statistically significant weighted mean effect sizes, regardless of the outcome measure. In addition, there was no statistically significant difference between retention (g = 0.63) and transfer (g = 0.53) outcomes. However, the retention-based outcome produced a much higher effect size, and was significantly larger than the mixed-retention and transfer outcome (g = 0.37).

How Are Effect Sizes Moderated by Contextual Features of the Research?

Table 4 shows the results of testing effects under different contextual features of research. Specifically, we investigated the degree to which the testing effect is moderated by retention interval, that is, the time between the practice and final tests. In addition, we examined the effects of initial learning criterion, specifically whether participants read passages, word lists, or word pairs. Table 4 shows that the majority of studies in our meta-analysis had a time lag of less than one day between the practice test and final test (k = 117 studies). The testing effect was robust across all intervals, suggesting that practice tests are effective, regardless of the time interval between the practice tests and the final test. Results also show that the highest weighted effect size was obtained with a time lag of 1 to 6 days (g = 0.82; k = 58). The between-levels difference was statistically significant, $Q_B(4) = 69.07$, $p < .001$. Post hoc analysis revealed that the use of practice tests was

TABLE 3

Weighted mean effect sizes for students and study characteristics

Moderator	*N*	*k*	Effect size		95% confidence interval		Test of heterogeneity		
			g+	*SE*	Lower	Upper	*Q*	Degrees of freedom	I^2 (%)
Study setting									
Classroom	2,722	30	0.67*	0.04	0.60	0.76	135.73*	29	79.37
Laboratory	11,687	223	0.62*	0.02	0.59	0.66	1169.61*	222	81.02
Not reported	1,018	19	0.44*	0.06	0.32	0.57	81.57*	18	77.93
Between-levels (Q_B)							18.09*	2	
Experimental design									
Random assignment	3,558	59	0.64*	0.04	0.57	0.71	371.65*	58	84.39
Nonrandom assignment	1,715	15	0.47*	0.05	0.38	0.57	62.25*	14	77.51
Within-subjects	3,988	82	0.60*	0.03	0.53	0.66	435.21*	81	81.39
Mixed	3,683	80	0.62*	0.03	0.55	0.68	301.23*	79	73.77
Not reported	2,483	36	0.70*	0.04	0.61	0.78	221.91*	35	84.23
Between-levels (Q_B)							12.75*	4	
Educational level									
Primary (Grades 1–6)	480	10	0.64*	0.10	0.45	0.83	51.50*	9	82.52

(continued)

TABLE 3 (continued)

Moderator	*N*	*k*	Effect size		95% confidence interval		Test of heterogeneity		
			g+	*SE*	Lower	Upper	*Q*	Degrees of freedom	I^2 (%)
Secondary (Grades 7–12)	1,335	19	0.83*	0.06	0.71	0.94	89.72*	18	79.94
Postsecondary	12,648	228	0.60*	0.02	0.57	0.64	1182.93*	227	80.81
Mixed (primary and postsecondary)	94	1	−0.65	0.21	−1.06	−0.24	0.00	0	0.00
Not reported	870	14	0.56*	0.07	0.43	0.70	30.35*	13	57.17
Between-levels (Q_B)							50.50*	4	
Outcome construct									
Retention	13,464	243	0.63*	0.02	0.60	0.67	1292.58*	242	81.28
Transfer	715	11	0.53*	0.08	0.38	0.68	36.14*	10	72.33
Mixed retention and transfer	1,038	15	0.37*	0.07	0.25	0.50	50.82*	14	72.45
Not reported	210	3	0.81*	0.15	0.51	1.10	8.00*	2	74.99
Between-levels (Q_B)							17.46*	3	

*$p < .05$.

TABLE 4

Weighted mean effect sizes for contextual features

Moderator	*N*	*k*	Effect size		95% confidence interval		Test of heterogeneity		
			g+	*SE*	Lower	Upper	*Q*	Degrees of freedom	I^2 (%)
Retention interval									
<1 day	5,766	117	0.56*	0.03	0.51	0.62	644.32*	116	82.00
1 day–6 days	2,978	58	0.82*	0.04	0.75	0.90	333.80*	57	82.92
7–42 days	3,667	52	0.69*	0.03	0.62	0.76	152.03*	51	66.45
Mixed	1,836	35	0.51*	0.05	0.42	0.61	158.57*	34	78.56
Not reported	1,180	10	0.30*	0.06	0.18	0.42	47.20*	9	80.93
Between-levels (Q_B)							69.07*	4	
Initial learning criterion									
Reading or studying a passage	4,143	66	0.71*	0.03	0.64	0.77	278.06*	65	76.62
Reading or studying word list	2,577	51	0.56*	0.04	0.48	0.64	484.22*	50	89.67
Reading or studying word pair/cue	4,635	99	0.63*	0.03	0.57	0.69	418.71*	98	76.59
Other	4,072	56	0.54*	0.03	0.47	0.60	208.69*	55	73.65
Between-levels (Q_B)							15.32*	3	
Time matched between tests/comparison group?									
No	2,357	43	1.00*	0.04	0.91	1.08	188.94*	42	77.77
Yes	8,867	163	0.50*	0.02	0.45	0.54	748.44*	162	78.36
Unclear	373	6	0.59*	0.11	0.38	0.81	17.95*	5	72.15
Each condition self-paced	2,198	37	0.57*	0.04	0.49	0.66	112.95*	36	68.13
Mixed (Yes and No)	1,140	18	0.68*	0.06	0.56	0.81	59.03*	17	71.20
Not applicable	492	5	1.35*	0.11	1.13	1.56	128.63*	4	96.89
Between-levels (Q_B)							149.04*	5	

*$p < .05$.

associated with a higher weighted mean effect size and differed significantly when the time lag between a practice test and a final test was between 1 and 6 days, in comparison to less than 1 day. We also found that although most studies ($k = 163$) matched the time between treatment and comparison groups, the largest effect size occurred when the time between treatment and comparison groups was not matched ($g = 1.00$), even though all categories produced statistically significant effect sizes.

In addition, Table 4 shows that participants benefited most from practice tests when the initial learning involved reading or studying a passage ($g = 0.71$). The between-levels difference was statistically significant, $Q_B(3) = 15.32$, $p < .001$. Post hoc analysis revealed that the use of practice tests was associated with higher weighted mean effect size when the initial learning involved reading or studying a passage and that the mean effect size is significantly different from results obtained when the initial learning involved reading or studying word lists.

How Are Effect Sizes Moderated by Methodological Features of the Research?

Table 5 shows the results of testing effects under varying methodological features of research. Specifically, we investigated the testing effects with reliability measures of the tests, fidelity of implementation (treatment fidelity), control for preexisting differences, and methodological quality of studies. A majority of the studies did not report the reliability of the tests used. These studies produced a moderately large weighted mean effect size ($g = 0.65$; $k = 242$). Studies that reported reliability measures were associated with a moderate weighted mean effect size ($g = 0.41$). Results in Table 5 also show the effects of practice tests when participants' preexisting knowledge was controlled. The weighted mean effect sizes were statistically significant across all forms of prior knowledge control. The largest effect size was obtained for studies using random assignment to control for preexisting differences in knowledge ($g = 0.72$). The between-levels difference was statistically significant, $Q_B(6) = 41.86$, $p < .001$. Post hoc analysis revealed that studies using a random assignment to control for preexisting differences in knowledge had a higher weighted mean effect size and differed significantly from studies using other means, including pretests and matching of participants.

Most studies (about 86%) were coded as *high* treatment fidelity, since implementation of treatment was monitored. Such high–treatment fidelity studies had a large weighted mean effect size ($g = 0.65$). Studies coded as low in fidelity of implementation also produced a statistically detectable moderate weighted mean effect size ($g = 0.46$). The between-levels difference was statistically significant, $Q_B(2) = 26.29$, $p < .001$. Post hoc analysis revealed that studies rated as high in treatment fidelity had a higher weighted mean effect size and differed significantly from studies that were rated as low in treatment fidelity. Finally, we coded for methodological quality. Studies that reported all information for extracting effect sizes, design, and monitored implementation of treatment were coded high in quality. Conversely, studies that did not sufficiently monitor implementation were coded as low in quality. The majority of the studies were coded as high quality (77%; $k = 208$) in terms of methodological quality. Beneficial effects of practice tests were found, regardless of the methodological quality of the studies.

TABLE 5

Weighted mean effect sizes for methodological features

Moderator	*N*	*k*	Effect size		95% confidence interval		Test of heterogeneity		
			g+	*SE*	Lower	Upper	*Q*	Degrees of freedom	I^2 (%)
Reliability									
Not reported	13,306	242	0.65*	0.02	0.61	0.69	1261.60*	241	80.90
Reported	1,963	26	0.41*	0.05	0.32	0.50	111.45*	25	77.57
Other	158	4	0.28	0.16	−0.03	0.59	4.26	3	29.53
Between-levels (Q_B)							27.69*	2	
Control for preexisting knowledge									
Not reported	3,336	49	0.58*	0.04	0.50	0.65	310.13*	48	84.52
Random assignment	3,988	75	0.72*	0.03	0.65	0.79	368.24*	74	79.90
Pretest/covariate	1,193	20	0.52*	0.06	0.40	0.63	107.87*	19	82.39
Matching of participants	285	5	0.36*	0.12	0.12	0.59	21.69*	4	81.56
Mixed	781	15	0.29*	0.07	0.15	0.43	15.10	14	7.26
Within-subjects	4,982	101	0.66*	0.03	0.60	0.72	524.95*	100	80.95
Other	862	7	0.55*	0.07	0.42	0.69	15.17*	6	60.44
Between-levels (Q_B)							41.86*	6	
Treatment fidelity									
High	12,978	236	0.65*	0.02	0.61	0.68	1265.14*	235	81.42
Low	2,329	34	0.46*	0.04	0.38	0.55	82.47*	33	59.99
Not reported	120	2	0.00	0.19	−0.38	0.38	31.10*	1	96.78
Between-levels (Q_B)							26.29*	2	
Methodological quality									
High	11,532	208	0.61*	0.02	0.57	0.64	1044.53*	207	80.18
Low	3,895	64	0.64*	0.03	0.58	0.71	359.58*	63	82.48
Between-levels (Q_B)							0.89	1	

*$p < .05$.

Publication Bias

All studies that met our inclusion criteria (except one) were published in peer-reviewed journals, heightening the potential for publication bias and limiting the possibility of investigating the potential moderating effect of the source of publication (published vs. unpublished). Studies with statistically significant results are published more often than studies with no significant results, which typically remain unpublished or at best reported in the less accessible grey literature (Lipsey & Wilson, 2001; Rosenthal, 1979). It is not surprising that aggregation of such statistically significant studies through a meta-analysis produces a significant overall mean effect size. However, this poses a considerable threat to the validity of meta-analytical results. To address this problem, meta-analysts recommend an examination of publication bias to ascertain validity of meta-analysis results.

Researchers have developed different statistical methods for estimating publication bias in meta-analyses. Indeed, there is some evidence that publication bias (or lack of it) is established through consistent results among the different methods (Ferguson, 2007; Rothstein, Sutton, & Borenstein, 2005). Therefore, we conducted two statistical tests using Comprehensive Meta-Analysis software to assess the potential for publication bias. First, we conducted a classic fail-safe *N* test to determine the number of null effect studies needed to raise the *p* value associated with the mean effect above an arbitrary alpha level ($\alpha = .05$). Results showed that 3,608 additional studies would be required to invalidate the overall effect. For the second test, we used Orwin's fail-safe *N* to estimate the number of file drawer studies with null results required to invalidate the overall effects found. Using a criterion trivial level of .05, results from the fail-safe *N* showed that 2,604 missing null studies are required to bring our current mean effect size to a trivial level of .05. Researchers claim that results from meta-analyses are valid and thus resistant to the file drawer problem if the fail-safe *N* reaches the $5k + 10$ limit (Carson, Schriesheim, & Kinicki, 1990; Rosenthal, 1979). Results of the two computed statistical tests suggest that publication bias does not pose a significant threat to the validity of our findings since both fail-safe *N* values are larger than the $5k + 10$ limit.

Discussion

The purpose of this meta-analysis was to summarize the learning benefits of taking a practice test versus other forms of non-testing learning conditions. Specifically, we examined all studies that compared learning benefits on a final test among practice retrieval and nonretrieval learning conditions. Results from 272 independent effects from 118 separate experiments indicate that testing effects across studies ($g = 0.61$) were robust, and that the strength of effect sizes was moderated by several variables. Findings from this meta-analysis provide a comprehensive understanding of the conditions under which practice tests enhance or inhibit learning, and offer empirical guidance to advance the theory of how learning processes are affected by the testing effect. In this section, we use our findings to provide evidence for and discuss why retrieval practice is an effective learning strategy. The discussion is organized around the research questions.

What Are the Learning Effects of Taking a Practice Test Compared With Other Learning Conditions?

As expected, testing effects were much stronger for comparison conditions in which participants did not perform any study activity or performed some filler activity unrelated to the final test ($g = 0.93$) compared to conditions using a non-retrieval study activity such as restudying or rereading ($g = 0.51$). Butler and Roediger (2007) point out that comparing a retrieval practice condition to a no-activity condition is a more realistic replication of a classroom setting. For example, a short quiz before a chapter exam would likely include some (but not all) questions seen on the chapter exam, and would likely not coincide with the restudy of other material on the exam (therefore replicating retrieval practice vs. no-activity conditions in a controlled laboratory experiment). However, when a practice test condition outperforms no-activity or filler conditions, one cannot assume testing effects occurred, because a more obvious explanation could be that the testing phase merely offered another opportunity to learn the material (Roediger & Karpicke, 2006a). Thus, the effect size of $g = 0.51$ is arguably the most accurate indicator of the benefits of retrieval practice, particularly for everyday application of students' study habits.

Empirical studies often use *rereading* as a comparison condition, in part because of its frequent use in real-world settings. In preparation for exams, students often reread their notes, lecture materials, and textbooks repeatedly. Indeed, survey research by Karpicke, Butler, and Roediger (2009) found that 84% of their sample of college students' use rereading as a strategy and that 55% reported rereading to be their most used study strategy. Unfortunately, the use of rereading as a study strategy may be minimally effective in light of Callender and McDaniel's (2009) finding that retention on a subsequent test was generally not enhanced by reading ecologically valid texts a second time. Rereading may be ineffective, in part, because it can give students a false sense of mastery, resulting in overconfidence and underpreparedness for exams. In other words, when students reread texts multiple times their *familiarity* of those texts increases, although this increase in familiarity does not necessarily mean that material has been *learned* sufficiently. Koriat and Bjork (2005) suggest that this overconfidence may result from a discord between what information is available during study sessions compared to what information is available during exams. Despite confidence in their understanding of study material when rereading it, students actually may not be able to answer exam questions related to the material. Nevertheless, our results provide support for the superiority of study strategies that incorporate retrieval over other nontesting study strategies.

Do Various Features of Practice Tests Foster Different Learning Benefits on a Final Test?

Test Format

Practice test format emerged as a significant moderator in this meta-analysis. Interestingly, differences between multiple-choice and short-answer practice test formats emerged ($g = 0.70$ and $g = 0.48$, respectively). Considering the relative ease of assessing multiple-choice tests compared with other forms of tests (e.g., free-recall,

short-answer), this result may encourage widespread use of multiple-choice tests for learning. However, one study showed that short-answer practice tests are more beneficial for long-term retention than multiple-choice tests (Kang et al., 2007). On the other hand, C. D. Morris, Bransford, and Franks's (1977) research on levels of processing suggests that retention is strongest when processing demands are *less* demanding. They reason that this is because less demanding retrieval practice activities allow participants to focus all of their cognitive energy on a simple task at hand, whereas deeper levels of processing require more cognitive energy and can distract participants from relevant aspects (C. D. Morris et al., 1977).

Another type of test format that requires relatively simple cognitive processes is cued-recall. Our data show that using retrieval to learn paired associates was much more effective than reading paired associates together ($g = 0.58$). This finding has strong implications for memorization of paired associates, especially in terms of language learning (e.g., recalling the Spanish equivalent of an English word) or geography lessons (e.g., attempting to recall the capitals of each U.S. state).

More important, this finding brings to mind the cognitive processes engaged in each question format. For example, we found that multiple-choice testing was the most effective format; however, this should be interpreted with caution, since an educator's decision to use any given format should be based on the content of the learning material and the expected learning outcomes. For example, multiple-choice tests may be especially useful for memorization and fact retention, while short-answer testing may require more higher order thinking skills that are useful for more conceptual and abstract learning content.

Since many studies use a combination of multiple-choice and short-answer questions during the practice phase, each requiring a different level of cognitive processing, the large effect size of $g = 0.80$ we obtained for mixed-format practice tests is relevant. The effectiveness of the mixed format may be due to interleaving, a technique that requires students to reload different cognitive processes. This forces learners to resolve the interference between different processes, resulting in long-term retention and transfer (E. L. Bjork & Bjork, 2011). In light of the large effect size, students may be encouraged to engage in practice testing using different testing formats. This suggestion and result may be more meaningful when considering the relevance of TAP.

Transfer-Appropriate Processing

We found that TAP moderated the strength of testing effects, indicating that testing effects were stronger when the practice and final test formats were identical ($g = 0.63$) than when the formats differed ($g = 0.53$). However, it is important to consider whether consistency between practice test and final test formats actually improves learning, or if this merely inflates test scores through familiarity and recognition. Recall (cued and free) was consistently chosen as the preferred test format by researchers for practice and final tests, and TAP was used in the design in about 64% of the studies coded. Studies investigating partial adherence to TAP, in which a combination of familiar and new test formats between practice and final tests was used, showed a large effect size of $g = 0.75$. However, this occurred in only about 10% of coded studies, highlighting the need for further research to understand the theoretical underpinnings of the TAP phenomenon and

its relationship with testing effects. The basic tenet of TAP suggests that memory traces are easiest to retrieve when retrieval processes are similar to how they were encoded during an initial learning activity. In this context, the act of retrieving the targeted information (answer) based on a cue (exam question) replicates the cognitive processes performed during a study activity that involves retrieval practice. Therefore, we coded TAP based on whether practice and final tests used the same format, moderating the strength of testing effects. However, testing effects themselves may be due to the use of TAP to some extent, since both the learning condition and outcome measure require the same cognitive processes (i.e., retrieval).

Number and Timing of Practice Tests

Our results show that it is more effective when students take a single practice test prior to the final test than when they take several practice tests. One hypothesis based on R. A. Bjork and Bjork's (1992) theoretical work is that recent retrieval practice activates relevant information in the working memory (Shea & Morgan, 1979). However, if we consider the time lapse between practice and final tests, we observe that interventions with a retention gap of less than 1 day had a smaller weighted effect size than those of 1 to 6 days ($g = 0.56$ and $g = 0.82$, respectively). Thus, our findings suggest that although a single test prior to a final test may result in better performance, the timing of the test should be carefully considered. One plausible explanation is more time between the practice and final tests allows students to mentally recall and process information, leading to deeper learning. An alternative hypothesis is that multiple tests within a short time may result in test fatigue that affects performance, while retrieval practice over a distributed time period enables long-term storage.

To What Degree Would the Testing Effect Vary When the Initial Practice Test Was Given With or Without Feedback?

An important variable that we coded was whether participants received feedback on their practice test before taking a final test. Findings indicate that a practice test followed by feedback or other forms of reexposure to the original study material did not yield higher testing effects than retrieval conditions that did not provide feedback. It is notable that reexposure (i.e., feedback or an additional learning opportunity after the practice test) was typically balanced in the studies we analyzed, meaning that students in both testing and nontesting conditions were reexposed to feedback or study material or neither was reexposed. The fact that feedback did not moderate testing effects indicates that reexposure to tested material does not necessarily strengthen final test scores *more* than reexposure to nontested material strengthens final test scores. In other words, testing effects were observed, regardless of whether feedback was provided. This suggests that students can be encouraged to use practice tests even when they do not receive feedback on such tests.

Our results do not imply that providing feedback on practice tests will or will not provide gains in retention. Instead, we suggest that a potential explanation for feedback not moderating the testing effect is that attempting to retrieve information from long-term memory is more cognitively challenging than study strategies students commonly use, such as passive reading of texts. Gardiner, Craik, and

Bleasdale (1973) suggest that the benefits of any given learning activity are strongly dependent on the cognitive effort required to perform that activity. Thus, we suggest that retrieval practice leads to deeper learning that may or may not be mitigated by the availability of feedback. This hypothesis is supported by Pyc and Rawson's (2009) finding that on altering the level of difficulty required for various retrieval attempts, testing effects became stronger as retrieval difficulty increased. It is notable that few studies analyzed reported information on the nature of feedback to investigate nuances. Hence, it is likely that feedback in certain situations may or may not contribute to improved learning performance.

It is surprising that feedback did not moderate the strength of testing effects, given the plethora of findings that testing *plus* feedback is more beneficial than testing without feedback in individual studies (e.g., Finley, Benjamin, Hays, Bjork, & Kornell, 2011; Metcalfe, Kornell, & Finn, 2009; Pashler, Cepeda, Wixted, & Rohrer 2005). In addition to our explanation above, unidentified between-groups confounds may have contributed to the null moderating effect of feedback. Also, we could not code for specific aspects of feedback (e.g., type of feedback; immediate versus delayed, etc.) due to an insufficient number of studies utilizing various feedback conditions. Future research should vary the conditions of feedback in testing effect studies to explore the types and features of feedback that most effectively enhance learning from practice tests.

How Do Testing Effects Vary When Used for Learning in Different Settings, Designs, and Educational Levels and With Different Outcome Constructs?

Study Setting

Testing effects remained consistently strong across settings ($g = 0.67$, $k = 30$, for classroom; $g = 0.62$, $k = 223$, for laboratory settings). This finding has especially strong implications for educational practice. Our findings indicate that retrieval practice improves retention on authentic academic exams more than nontesting control conditions do. However, other factors were not balanced across this factor, which could confound or moderate results. We noticed several methodological differences between classroom-based and laboratory-based studies. First, the retention interval between practice and final tests was longer in classrooms. Of those studies coded for that reported retention interval, 67% of classroom studies reported retention intervals between 7 and 42 days, while 59% of laboratory studies had a retention interval of less than 1 day. Furthermore, one or more quizzes are often administered throughout the semester followed by a final exam at the end of the term in classroom-based studies. In contrast, laboratory studies are often conducted during one testing session, with very short retention intervals, sometimes by separating tests with short filler activities.

A second difference between study settings is the formats of practice and final tests. In laboratory settings, free- or cued-recall testing procedures were often used (75% in practice tests and 62% in final tests), whereas in classroom settings, multiple-choice–only (37% in practice tests and 47% in final tests) and mixed formats (23% in practice tests) were more popular. Finally, although a majority of classroom- and laboratory-based studies focused on postsecondary students, there were more classroom-based studies (33%) administered in K–12 settings than laboratory studies (5%). Meta-regression analyses would be useful for exploring

the confounding effects of the relationships between study setting and other factors described here. In light of these potential confounds, *comparison* of classroom and laboratory effect sizes should be interpreted with caution. Regardless of methodological differences, it is notable that strong testing effects emerge for both laboratory and classroom settings.

Education Level

Although about 83% of the studies we analyzed used samples of postsecondary students, a substantial amount used samples of primary or secondary students. Our findings indicate that testing effects are strongest for secondary students (g = 0.83), outperforming primary (g = 0.64) and postsecondary students (g = 0.60). More important, results show that retrieval practice appears to produce robust testing effects across all educational levels, indicating that retrieval practice should be incorporated into a wide array of educational settings. That said, nearly all studies reviewed in this meta-analysis were conducted with participants without cognitive or intellectual impairment (cf. Sumowski, Chiaravalloti, & DeLuca, 2010; Sumowski, Wood, et al., 2010). Therefore, future research may explore the benefits of retrieval practice in special education classrooms with atypical populations, such as students with learning disabilities or developmental disorders.

Learning Construct

We found that the testing effect was always present, regardless of outcome construct. However, only 11 studies examined testing effects on transfer outcome measures. Future research may examine the robustness of testing effects with transfer measures.

How Are Effect Sizes Moderated by Contextual Features of the Research?

Experimental design emerged as a significant moderator. For example, randomized designs (g = 0.64) outperformed nonrandomized or quasi-experimental designs (g = 0.47). This important result suggests that effects of practice tests are apparent even with more stringent designs that use random assignment. Quasi-experimental designs are often used in classroom-based studies, because random assignment within a classroom would create ethical and logistical challenges for instructors, and different students in the same classroom may experience different modes of instruction or learning activities. Instead, researchers sometimes compared two separate classrooms in which one used practice tests and the other did not. Our results also show no significant difference between within-subjects and between-subjects designs, indicating that within-subjects may be a useful matching procedure for classrooms in which randomization is not possible.

Practical Implications

Previous studies show that most students rely on rereading textbooks, notes, and other study materials (Karpicke et al., 2009), which has been found to be ineffective (Callender & McDaniel, 2009). Although Roediger and Karpicke (2006b) have advocated for "test-enhanced learning" in mainstream classrooms, it is not surprising that their message is often met with resistance, given the controversy and negative connotations of testing. Indeed, in a review of three policy-oriented journals,

Buck, Ritter, Jensen, and Rose (2010) found that 90% of articles were critical of testing. As standardized testing has skyrocketed in recent years, educators may be understandably opposed to more testing. In agreement with Roediger and Karpicke (2006b), we advocate for the use of frequent *low-stakes* quizzes, as a learning tool so that teachers and students can assess knowledge gaps, rather than high-stakes tests used only for summative purposes and high-stakes decision-making.

For better or worse, the current educational climate of K–12 education requires students to meet measurable benchmarks on several standardized tests throughout their schooling. We suggest that the use of retrieval practice learning activities will help students develop test-taking skills that may improve performance on high-stakes tests. Although more research is needed on transfer-based outcomes, our results show that practice tests provide similar testing effects for both retention and transfer-based outcomes. Transfer, some would argue, is the ultimate goal of education (Bransford, Brown, & Cocking, 1999) so that knowledge and thinking abilities learned in schools can be transferred to a range of real-world settings. Practice tests should be constructed to promote transferrable, higher order thinking skills. Instructors may use well-crafted multiple-choice and short-answer tests to gauge students' prior knowledge before instruction begins. This can help teachers identify students' misconceptions and plan instructions to correct them.

Our recommendation for more classroom testing may at first be misconstrued. While we maintain that increased formative practice testing is a good idea, students can benefit from retrieval practice in many other ways, without suffering through more summative tests. Indeed, retrieval practice need not come in the form of a quiz, and can easily be incorporated during self-directed studying (i.e., flash cards or self-generated questions), during structured learning activities in the classroom, or even lectures. For example, Tobin (1987) reviewed considerable evidence demonstrating that teacher *wait time*—3- to 5-second pause separating utterances during verbal interaction—improves achievement and facilitates higher level learning by providing students with additional time to think. The same benefits can be skillfully applied during classroom lectures in which teachers stop to ask questions. Instead of immediately calling on the first student to raise their hand, teachers may pause for several seconds to let students *think* and generate answers (i.e., retrieval practice). This simple strategy allows all students in class to benefit *cognitively* (through the act of retrieval) and *metacognitively* (by assessing how well they knew the answer to the question after it has been answered).

Comprehension monitoring, a metacognitive skill known also as "metacomprehension," is a particularly important outcome of quizzing (especially during self-directed study activities). As Dunlosky and Lipko's (2007) review of their own extensive research programs showed, most students are remarkably poor at judging whether or not they have studied a piece of material well enough to have mastered it. This skill is called "judgments of learning" (JOLs). Research on JOLs demonstrates that retrieval practice substantially improves JOLs, since the accurate or inaccurate retrieval of information is a clear indication to the learner regarding mastery. Students can use flash cards and quizzes to benefit from retrieval and evaluate which topic areas need increased study effort. In this way, retrieval practice can be a learning activity in itself or seamlessly incorporated into regular classroom activities.

The majority of research reviewed in this meta-analysis has focused on university undergraduates in both laboratory and classroom settings. Perhaps there are specific reasons why university classrooms may benefit from increased retrieval practice. Leeming (2002) demonstrated that beside improvements in overall course grades, frequent classroom quizzing increases class attendance. Students subjectively rated courses as more enjoyable and beneficial than the same courses with no frequent classroom quizzing—a finding corroborated by Bangert-Drowns et al.'s (1991) review. Although instructors may hesitate to administer quizzes due to the burden of grading, quick-response technologies (e.g., clickers) and learning management systems such as *Canvas*, *Moodle*, *Blackboard*, and *WebCT* often have built-in quiz features that can be incorporated into gradebooks with minimal administrative effort. For example, clicker technology is becoming increasingly common in classrooms, as it (a) allows instructors to examine mastery of knowledge during class, (b) allows students to become actively engaged in large classrooms where student-teacher interaction is otherwise minimal, (c) provides increased incentive for students to attend class, and (d) provides immediate feedback to students and teachers in order to monitor comprehension (Hunsu, Adesope, & Bayly, 2016; Mayer et al., 2009).

Limitations and Future Research Directions

Educators who aim to incorporate retrieval practice into structured class time should be cognizant of how low-stakes quizzing may relate to the learning goals of their curriculum. They should be cautious not to use quizzing as a means to "teach to the test." It is important to emphasize that the goal of education is to promote meaningful learning, not to inflate test scores through recognition and rote memory. To demonstrate that retrieval practice does the former and not the latter, researchers must show that learning gains through practice tests can be observed in subsequent tests using alternate test items and content to demonstrate transfer-based learning outcomes. Some research has shown promising results (e.g., Butler, 2010), but more research is needed in this area. We are also aware of the negative associations and possible anxiety associated with testing, which may restrict implementation of our findings. Therefore, it is important to reiterate that this meta-analysis presents results from low-stakes practice tests, not high-stakes testing, and pays careful attention to examining the robustness of findings in natural school contexts (i.e., in the classroom). To improve the generalizability and ecological validity of testing effects, future research should strategically examine whether decades of findings from the laboratory transfers to actual classroom settings. We encourage researchers to examine the effects and interplay of testing on imposed social, political, and behavioral factors using different methods to develop a deeper understanding of the testing phenomenon.

Another area for future research is to explore individual differences that may influence testing effects. For example, Callender and McDaniel (2007) demonstrated that testing effects are stronger for participants with low reading comprehension abilities. Sumowski, Chiaravalloti, et al. (2010) demonstrated that individuals with multiple sclerosis (a neurological disease associated with memory dysfunction), benefit *less* than neurologically normal controls from various

nonretrieval learning conditions; however, they benefit equally from retrieval practice. Although this suggests that students with learning disabilities or other cognitive abnormalities may benefit from retrieval practice, more research is needed to investigate specific populations and cognitive profiles that may benefit from retrieval, since the vast majority of research to date has been conducted on cognitively healthy undergraduates.

In addition, future research may explore and document the effects of various feedback variables, including the type, speed, and medium of feedback provided when investigating the testing effect. It may also be useful to examine differences in testing effects based on domain or content used in the studies or mode of delivery of learning intervention. Furthermore, given the plethora of studies comparing retrieval practice to rereading, future research may robustly examine the benefits of retrieval practice compared with other effective, active learning techniques such as elaborative interrogation and self-explanation (see Roediger & Pyc, 2012).

Finally, we acknowledge potential concerns on the possible lack of representativeness of included studies, since the Dissertation Abstracts database was not searched. However, we searched the PsychINFO database, which also houses several dissertations. Indeed, about 12% of all studies archived in the PsychINFO database consists of dissertations abstracted from Dissertation Abstracts International (American Psychological Association, 2016). To further address this limitation, we conducted sensitivity analyses and found that across two different analyses, over 2,600 additional studies are required to invalidate the overall effect found in this meta-analysis. The sensitivity analyses suggest that findings from this meta-analysis are unlikely to have suffered from publication bias. Nevertheless, the noninclusion of unpublished studies remains a limitation in the present meta-analysis and should be addressed in future studies.

Conclusion

An overwhelming amount of evidence reviewed in this meta-analysis suggests that retrieval practice increases achievement. The benefits of retrieval practice persist across a wide array of educational levels, settings, and testing formats and procedures. Therefore, students should be encouraged and taught how to use retrieval practice during self-directed learning activities, and teachers may incorporate retrieval practice into structured classroom activities. In sum, results of this systematic and evidence-based meta-analysis provide a platform to help educators rethink other ways in which tests and other forms of retrieval practice could be used to promote learning. Once stakeholders realize the cognitive, metacognitive, and noncognitive benefits of practice tests, rather than only using summative assessments for high-stakes decisions, findings of this evidence-based research may be used to inform educational practice in K–12 and tertiary settings.

Note

This research was supported by Washington State University's College of Education Faculty Funding Award to Olusola O. Adesope.

References

References marked with an asterisk indicate studies included in the meta-analysis.

Adesope, O. O., Lavin, T., Thompson, T., & Ungerleider, C. (2010). A systematic review and meta-analysis of the cognitive correlates of bilingualism. *Review of Educational Research*, *80*, 207–245. doi:10.3102/0034654310368803

Adesope, O. O., & Nesbit, J. C. (2012). Verbal redundancy in multimedia learning environments: A Meta-Analysis. *Journal of Educational Psychology*, *104*, 250–263. doi:10.1037/a0026147

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

Ahn, S., Ames, A. J., & Myers, N. D. (2012). A review of meta-analyses in education: Methodological strengths and weaknesses. *Review of Educational Research*, *82*, 436–476. doi:10.3102/0034654312458162

American Psychological Association. (2016, May 21). *A world-class resource for behavioral and social science research*. Retrieved from http://www.apa.org/pubs/databases/psycinfo/psycinfo-printable-fact-sheet.pdf

Anderson, J. R., Reder, L. M., & Lebiere, C. (1996). Working memory: Activation limitations on retrieval. *Cognitive Psychology*, *30*, 221–256. doi:10.1006/cogp.1996.0007

Anderson, M. C., Bjork, R. A., & Bjork, E. L. (1994). Remembering can cause forgetting: Retrieval dynamics in long–term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 1063–1087. doi:10.1037/0278–7393.20.5.1063

*Arnold, K., & McDermott, K. (2013). Free recall enhances subsequent learning. *Psychonomic Bulletin & Review*, *20*, 507–513. doi:10.3758/s13423-012-0370-3

Bangert–Drowns, R. L., Kulik, J. A., & Kulik, C. L. C. (1991). Effects of frequent classroom testing. *Journal of Educational Research*, *85*, 89–99. doi:10.1080/00220671.1991.10702818

*Barcroft, J. (2007). Effect of opportunities for word retrieval during second language vocabulary learning. *Language Learning*, *57*, 35–56. doi:10.1111/j.1467-9922.2007.00398.x

*Bishara, A. J., & Jacoby, L. (2008). Aging, spaced retrieval, and inflexible memory performance. *Psychonomic Bulletin & Review*, *15*, 52–57. doi:10.3758/PBR.15.1.52

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher, R. W. Pew, L. M. Hough, & J. R. Pomerantz (Eds.), *Psychology and the real world: Essays illustrating fundamental contributions to society* (pp. 56–64). Washington, DC: FABBS Foundation.

*Bjork, E. L., & Storm, B. C. (2011). Retrieval experience as a modifier of future encoding: another test effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *5*, 1113–1124. doi:10.1037/a0023549

Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In A. Healy, S. Kosslyn, & R. Shiffrin (Eds.), *From learning processes to cognitive processes: Essays in honor of William K. Estes* (Vol. 2, pp. 35–67). Hillsdale, NJ: Erlbaum.

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

Black, P., & Wiliam, D (1998). Assessment and classroom learning. *Assessment in Education*, *5*, 7–71. doi:10.1080/0969595980050102

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. R. (2008). *Comprehensive meta-analysis* (Version 2. 2.048) [Computer software]. Englewood, NJ: Biostat.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, England: Wiley.

*Bouwmeester, S., & Verkoeijen, P. P. J. L. (2011). The effect of instruction method and relearning on Dutch spelling performance of third- through fifth-graders. *European Journal of Psychology of Education*, *1*, 61–74. doi:10.1007/s10212–010–0036–3

Bransford, J. D., Brown, A. L., & Cocking, R. R. (1999). *How people learn* (J. D. Bransford, Ed.). Washington, DC: National Academy Press.

Bransford, J. D., Franks, J. J., Morris, C. D., & Stein, B. S. (1979). Some general constraints on learning and memory research. In L. S. Cermak & F. I. M. Craik (Eds.), *Levels of processing in human memory* (pp. 331–354). Hillsdale, NJ: Erlbaum.

Buck, S., Ritter, G. W., Jensen, N. C., & Rose, C. P. (2010). Teachers say the most interesting things: An alternative view of testing. *Phi Delta Kappan*, *91*, 50–54. doi:10.1177/003172171009100613

*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. doi:10.1037/a0019902

*Butler, A. C., Karpicke, J. D., & Roediger, H. L. (2007). The effect of type and timing of feedback on learning from multiple-choice tests. *Journal of Experimental Psychology: Applied*, *13*, 273–281. doi:10.1037/1076–898X.13.4.273

*Butler, A. C., Karpicke, J. D., & Roediger, H. L. (2008). Correcting a metacognitive error: Feedback enhances retention of low–confidence correct responses. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 918–928. doi:10.1037/0278–7393.34.4.918

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

Butler, A. C., & Roediger, H. L. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition*, *36*, 604–616. doi:10.3758/MC.36.3.604

*Callender, A. A., & McDaniel, M. A. (2007). The benefits of embedded question adjuncts for low and high structure builders. *Journal of Educational Psychology*, *99*, 339–348. doi:10.1037/0022–0663.99.2.339

Callender, A. A., & McDaniel, M. A. (2009). The limited benefits of rereading educational texts. *Contemporary Educational Psychology*, *34*, 30–41. doi: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. doi:10.1037/a0017021

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

*Carpenter, S. K., & DeLosh, E. L. (2005). Application of the testing and spacing effects to name learning. *Applied Cognitive Psychology*, *19*, 619–636. doi:10.1002/acp.1101

*Carpenter, S. K., & DeLosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory & Cognition*, *34*, 268–276. doi:10.3758/BF03193405

*Carpenter, S. K., & Pashler, H. (2007). Testing beyond words: Using tests to enhance visuospatial map learning. *Psychonomic Bulletin & Review*, *14*, 474–478. doi:10.3758/BF03194092

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

*Carpenter, S. K., Pashler, H., & Vul, E. (2006). What types of learning are enhanced by a cued recall test? *Psychonomic Bulletin & Review*, *13*, 826–830. doi:10.3758/BF03194004

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

Carson, K. P., Schriesheim, C. A., & Kinicki, A. J. (1990). The usefulness of the "fail-safe" statistic (*N*) in meta–analysis. *Educational and Psychological Measurement*, *50*, 233–243. doi:10.1177/0013164490502001

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. doi:10.1016/j.jml.2009.04.004

*Chan, J. C. K. (2010). Long term effects of testing on the recall of nontested materials. *Memory*, *18*, 49–57. doi:10.1080/09658210903405737

*Chan, J. C. K., & McDermott, K. B. (2007). The testing effect in recognition memory: A dual process account. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*, 431–437. doi:10.1037/0278-7393.33.2.431

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

*Coane, J. (2013). Retrieval practice and elaborative encoding benefit memory in younger and older adults. *Journal of Applied Research in Memory and Cognition*, *2*, 95–100.doi:10.1016/j.jarmac.2013.04.001

Coburn, C. E., Hill, H. C., & Spillane, J. P. (2016). Alignment and accountability in policy design and implementation: The Common Core State Standards and implementation research. *Educational Researcher*, *45*, 243–251. doi:10.3102/0013189X16651080

*Congleton, A., & Rajaram, S. (2012). The origin of the interaction between learning method and delay in the testing effect: The roles of processing and conceptual retrieval organization. *Memory & Cognition*, *40*, 528–539. doi:10.3758/s13421-011-0168-y.

Cooper, H. M., Hedges, L. V., & Valentine, J. (Eds.). (2009). *The handbook of research synthesis and meta-analysis* (2nd ed.). New York, NY: Russell Sage Foundation.

*Coppens, L. C., Verkoeijen, P., & Rikers, R. (2011). Learning adinkra symbols: The effect of testing. *Journal of Cognitive Psychology*, *23*, 351–357. doi:10.1080/20445911.2011.507188

*Cranney, J., Ahn, M., McKinnon, R., Morris, S., & Watts, K. (2009). The testing effect, collaborative learning, and retrieval-induced facilitation in a classroom setting. *European Journal of Cognitive Psychology*, *21*, 919–940. doi:10.1080/09541440802413505

*Cull, W. L. (2000). Untangling the benefits of multiple study opportunities and repeated testing for cued recall. *Applied Cognitive Psychology*, *14*, 215–235. doi:10.1002/(SICI)1099-0720(200005/06)14:3<215::AID-ACP640>3.3.CO;2-T

*deJonge, M., & Tabbers, H. (2013). Repeated testing, item selection and relearning: The benefits of testing outweigh the costs. *Experimental Psychology*, *60*, 206–212. doi:10.1027/1618-3169/a000189

Delaney, P. F., Verkoeijen, P. P. J. L., & Spirgel, A. (2010). Spacing and testing effects: A deeply critical, lengthy, and at times discursive review of the literature. In B. H. Ross (Ed.), *Psychology of learning and motivation: Advances in research and theory* (Vol. 53, pp. 63–147). San Diego, CA: Academic Press.

*Duchastel, P. (1979). Retention of prose materials: The effect of testing. *Journal of Educational Research*, *72*, 299–300. doi:10.1080/00220671.1979.10885176

*Duchastel, P. (1980a). Effect of testing on retention of prose: A replication. *Psychological Reports*, *46*, 182. doi:10.2466/pr0.1980.46.1.182

*Duchastel, P. (1980b). Extension of testing effect on the retention of prose. *Psychological Reports*, *47*, 1062. doi:10.2466/pr0.1980.47.3f.1062

*Duchastel, P. (1981). Retention of prose following testing with different types of test. *Contemporary Educational Psychology*, *6*, 217–226. doi:10.1016/0361-476X(81)90002-3

*Duchastel, P. C., & Nungester, R. J. (1981). Long-term retention of prose following testing. *Psychological Reports*, *49*, 470. doi:10.2466/pr0.1981.49.2.470

Dunlosky, J., & Lipko, A. R. (2007). Metacomprehension: A brief history and how to improve its accuracy. *Current Directions in Psychological Science*, *16*, 228–232. doi:10.1111/j.1467-8721.2007.00509.x

*Einstein, G., Mullet, H., & Harrison, T. (2012). The testing effect: Illustrating a fundamental concept and changing study strategies. *Teaching of Psychology*, *39*, 190–193. doi:10.1177/0098628312450432

Executive Office of the President. (2015). *Every Student Succeeds Act. A progress report on elementary and secondary education* (White House Report). Retrieved fromhttps://www.whitehouse.gov/sites/whitehouse.gov/files/documents/ESSA_Progress_Report.pdf

Ferguson, C. J. (2007). Evidence for publication bias in video game violence effects literature: A meta-analytic review. *Aggression and Violent Behavior*, *12*, 470–482. doi:10.1016/j.avb.2007.01.001

*Finley, J. R., Benjamin, A. S., Hays, M. J., Bjork, R. A., & Kornell, N. (2011). Benefits of accumulating versus diminishing cues in recall. *Journal of Memory and Language*, *64*, 289–298. doi:10.1016/j.jml.2011.01.006

Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2006). *How to design and evaluate research in education* (Vol. 7). New York, NY: McGraw-Hill.

*Fritz, C. O., Morris, P. E., Nolan, D., & Singleton, J. (2007). Expanding retrieval practice: An effective aid to preschool children's learning. *Quarterly Journal of Experimental Psychology*, *60*, 991–1004. doi:10.1080/17470210600823595

Gardiner, J. M., Craik, E. I. M., & Bleasdale, E. A. (1973). Retrieval difficulty and subsequent recall. *Memory & Cognition*, *1*, 213–216.

Gates, A. I. (1917). Recitation as a factor in memorizing. *Archives of Psychology*, *6*, 1–104.

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

*Glover, J. A., Krug, D., Hannon, S., & Shine, A. (1990). The "testing" effect and restricted retrieval rehearsal. *Psychological Record*, *40*, 215–226.

*Golding, J. M., Wasarhaley, N. E., & Fletcher, B. (2012). The use of flashcards in an introduction to psychology class. *Teaching of Psychology*, *39*, 199–202. doi:10.1177/0098628312450436

*Grimaldi, P. J., & Karpicke, J. D. (2012). When and why do retrieval attempts enhance subsequent encoding? *Memory & Cognition*, *40*, 505–513. doi:10.3758/s13421-011-0174-0

*Halamish, V., & Bjork, R. A. (2011). When does testing enhance retention? A distribution-based interpretation of retrieval as a memory modifier. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*, 801–812. doi:10.1037/a0023219

*Hanawalt, N. G., & Tarr, A. G (1961). The effect of recall upon recognition. *Journal of Experimental Psychology*, *62*, 361–367. doi:10.1037/h0041917

Harwell, M., & Maeda, Y. (2008). Deficiencies of reporting in meta-analyses and some remedies. *The Journal of Experimental Education*, *76*, 403–430. doi:10.3200/JEXE.76.4.403-430

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*, 81–112. doi:10.3102/003465430298487

Hays, M. J., Kornell, N., & Bjork, R. A. (2010). Costs and benefits of feedback during learning. *Psychonomic Bulletin & Review*, *17*, 797–801. doi:10.3758/PBR.17.6.797

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational and Behavioral Statistics*, *6*, 107–128. doi:10.3102/10769986006002107

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. San Diego, CA: Academic Press.

Hedges, L. V., & Vevea, J. L. (1998). Fixed and random–effects models in meta–analysis. *Psychological Methods*, *3*, 486–504. doi:10.1037/1082–989X.3.4.486

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

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

Hunsu, N. J., Adesope, O., & Bayly, D. J. (2016). A meta-analysis of the effects of audience response systems (clicker-based technologies) on cognition and affect. *Computers & Education*, *94*, 102–119. doi:10.1016/j.compedu.2015.11.013

*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. doi:10.1037/a0020636

*Johnson, C. I., & Mayer, R. E. (2009). A testing effect with multimedia learning. *Journal of Educational Psychology*, *101*, 621–629. doi:10.1037/a0015183

*Jonsonn, F., Hedner, M., & Olsson, M. (2012). The testing effect as a function of explicit testing instructions and judgments of learning. *Experimental Psychology*, *59*, 251–257. doi:10.1027/1618-3169/a000150

*Jonsonn, F., Kubik, V., Sundqvist, M., Todorov, I., & Jonsson, B. (2013). How crucial is the response format for the testing effect? *Psychological Research*, *78*, 623–633. doi:10.1007/s00426-013-0522-8

*Kang, S. H. K. (2010). Enhancing visuo-spatial learning: The benefit of retrieval practice. *Memory & Cognition*, *38*, 1009–1017. doi:10.3758/MC.38.8.1009

*Kang, S. H. K., Gollan, T. H., & Pashler, H. (2013). Don't just repeat after me: Retrieval practice is better than imitation for foreign vocabulary learning.

Psychonomic Bulletin & Review, *20*, 1259–1265. doi:10.3758/s13423-013-0450-z

*Kang, S. H. K., McDermott, K. B., & Roediger, H. L. (2007). Test format and corrective feedback modify the effect of testing on long–term retention. *European Journal of Cognitive Psychology*, *19*, 528–558. doi: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. doi: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. doi:10.1126/science.1199327

Karpicke, J. D., Butler, A. C., & Roediger, H. L., III. (2009). Metacognitive strategies in student learning: Do students practise retrieval when they study on their own? *Memory*, *17*, 471–479. doi:10.1080/09658210802647009

*Karpicke, J. D., & Smith, M. A. (2012). Separate mnemonic effects of retrieval practice and elaborative encoding. *Journal of Memory and Language*, *67*, 17–29. doi:10.1016/j.jml.2012.02.004

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

*Karpicke, J. D., & Zaromb, F. M. (2010). Retrieval mode distinguishes the testing effect from the generation effect. *Journal of Memory and Language*, *62*, 227–239. doi:10.1016/j.jml.2009.11.010

*King, A. (1992). Comparison of self-questioning, summarizing, and notetaking-review as strategies for learning from lectures. *American Educational Research Journal*, *29*, 303–323. doi:10.3102/00028312029002303

Koriat, A., & Bjork, R. A. (2005). Illusions of competence in monitoring one's knowledge during study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 187–194. doi:10.1037/0278–7393.31.2.187

Kornell, N., & Bjork, R. A. (2007). The promise and perils of self–regulated study. *Psychonomic Bulletin & Review*, *6*, 219–224. doi:10.3758/BF03194055

*Kornell, N., Hays, M. J., & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 989–998. doi:10.1037/a0015729

*Kornell, N., & Son, L. K. (2009). Learners' choices and beliefs about self-testing. *Memory*, *17*, 493–501. doi:10.1080/09658210902832915

*Kromann, C. B., Jensen, M. L., & Ringsted, C. (2010). The testing effect on skills might last 6 months. *Advances in Health Sciences Education*, *15*, 395–401. doi:10.1111/j.1365-2923.2008.03245.x

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

Kulik, J. A., & Kulik, C. C. (1988). Timing of feedback and verbal learning. *Review of Educational Research*, *58*, 79–97. doi:10.3102/00346543058001079

*Lambert, T., & Saville, B. K. (2012). Interteaching and the testing effect: A preliminary analysis. *Teaching of Psychology*, *39*, 194–198. doi:10.1177/0098628312450435

*LaPorte, R. E., & Voss, J. F. (1975). Retention of prose materials as a function of postacquisition testing. *Journal of Educational Psychology*, *67*, 259–266. doi:10.1037/h0076933

*Larsen, D. P., Butler, A. C., Lawson, A. L., & Roediger, H. L., III. (2012). The importance of seeing the patient: test-enhanced learning with standardized patients and

written tests improves clinical application of knowledge. *Advances in Science Education*, *18*, 409–425. doi:10.1007/s10459-012-9379-7

*Larsen, D. P., Butler, A. C., & Roediger, H. L., III. (2009a). Comparative effects of test-enhanced learning and self-explanation on long-term retention. *Medical Education*, *47*, 674–682. doi:10.1111/medu.12141

*Larsen, D. P., Butler, A. C., & Roediger, H. L., III. (2009b). Repeated testing improves long-term retention relative to repeated study: A randomized, controlled trial. *Medical Education*, *43*, 1174–1181. doi:10.1111/j.1365-2923.2009.03518.x

*Leeming, F. C. (2002). The exam-a-day procedure improves performance in psychology classes. *Teaching of Psychology*, *29*, 210–212. doi:10.1207/S15328023TOP2903_06

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis: Applied social research methods series* (Vol. 49). Thousand Oaks, CA: Sage.

*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. doi:10.1177/0098628311401587

*Marsh, E. J., Agarwal, P. K., & Roediger, H. L. (2009). Memorial consequences of answering SAT II questions. *Journal of Experimental Psychology: Applies*, *15*, 1–11. doi:10.1037/a0014721

*Marsh, E. J., Fazio, L. K., & Goswick, A. E. (2012). Memorial consequences of testing school-aged children. *Memory*, *20*, 899–906. doi:10.1080/09658211.2012.708757

Marsh, E. J., Roediger, H. L., Bjork, R. A., & Bjork, E. L. (2007). The memorial consequences of multiple-choice testing. *Psychonomic Bulletin & Review*, *6*, 194–199. doi:10.3758/BF03194051

*Masson, M. E., & McDaniel, M. A. (1981). The role of organizational processes in long-term retention. *Journal of Experimental Psychology: Human Learning and Memory*, *7*, 100–110

*Mayer, R. E., Stull, A., DeLeeuw, K., Ameroth, K., Bimber, B., Chun, D., . . . Zhang, H. (2009). Clickers in college classrooms: Fostering learning with questioning methods in large lecture classes. *Contemporary Educational Psychology*, *3*, 51–57. doi:10.1016/j.cedpsych.2008.04.002

*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. doi:10.1037/a0021782

*McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology*, *19*, 494–513. doi:10.1080/09541440701326154

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

*McDaniel, M. A., Howard, D. C., & Einstein, G. O. (2009). The read–recite–review study strategy: Effective and portable. *Psychological Science*, *20*, 516–522. doi:10.1111/j.1467-9280.2009.02325.x

*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.doi:10.1002/acp.2914

*McDaniel, M. A., Wildman, K. M., & Anderson, J. L. (2012). Using quizzes to enhance summative-assessment performance in a web-based class: An experimental

study. *Journal of Applied Research in Memory and Cognition*, *1*, 18–26. doi:10.1016/j.jarmac.2011.10.001

*McDermott, K. B. (2006). Paradoxical effects of testing: Repeated retrieval attempts enhance the likelihood of later accurate and false recall. *Memory & Cognition*, *34*, 261–267. doi:10.3758/BF03193404

*Metcalfe, J., & Kornell, N. (2007). Principles of cognitive science in education: The effects of generation, errors and feedback. *Psychonomic Bulletin & Review*, *14*, 225–229. doi:10.3758/BF03194056

Metcalfe, J., Kornell, N., & Finn, B. (2009). Delayed versus immediate feedback in children's and adults' vocabulary learning. *Memory & Cognition*, *37*, 1077–1087. doi:10.3758/MC.37.8.1077

*Meyer, A. N. D., & Logan, J. M. (2013). Taking the testing effect beyond the college freshman: Benefits for lifelong learning. *Psychology and Aging*, *28*, 142–147. doi:10.1037/a0030890

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior*, *16*, 519–533. doi:10.1016/s0022–5371(77)80016–9

*Morris, P. E., & Fritz, C. O. (2002). The improved name game: better use of expanding retrieval practice. *Memory*, *10*, 259–266. doi:10.1080/09658210143000371

*Morris, P. E., Fritz, C. O., Jackson, L., Nichol, E., & Roberts, E. (2005). Strategies for learning proper names: Expanding retrieval practice, meaning and imagery. *Applied Cognitive Psychology*, *19*, 779–798. doi:10.1002/acp.1115.

Murayama, K., Miyatsu, T., Buchli, D., & Storm, B. C. (2014). Forgetting as a consequence of retrieval: A meta-analytic review of retrieval-induced forgetting. *Psychological Bulletin*, *140*, 1383–1409. doi:10.1037/a0037505

Nesbit, J. C., & Adesope, O. O. (2006). Learning with concept and knowledge maps: A meta-analysis. *Review of Educational Research*, *76*, 413–448. doi:10.3102/00346543076003413

*Nungester, R. J., & Duchastel, P. C. (1982). Testing versus review: Effects on retention. *Journal of Applied Psychology*, *74*, 18–22. doi:10.1037/0022-0663.74.1.18.

Pashler, H., Cepeda, N. J., Wixted, J. T., & Rohrer, D. (2005). When does feedback facilitate learning of words? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 3–8. doi:10.1037/0278–7393.31.1.3

*Pastötter, B., Schicker, S., Niedernhuber, J., & Bäuml, K. H. T. (2011). Retrieval during learning facilitates subsequent memory encoding. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*, 287–297. doi:10.1037/a0021801

*Peterson, D. J., & Mulligan, N. W. (2013). The negative testing effect and multifactor account. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *39*, 1287–1293. doi:10.1037/a0031337

Phelps, R. P. (2012). The effect of testing on student achievement, 1910–2010. *International Journal of Testing*, *12*, 21–43. doi:10.1080/15305058.2011.602920

*Pressley, M., Tanenbaum, R., McDaniel, M. A., & Wood, E. (1990). What happens when university students try to answer prequestions that accompany textbook material? *Contemporary Educational Psychology*, *15*, 27–35. doi:10.1016/0361-476X(90)90003-J

*Pu, X., & Tse, C. (2013). The influence of intentional versus incidental retrieval practices on the role of recollection in test-enhanced learning. *Cognitive Processing*, *15*, 55–64. doi:10.1007/s10339-013-0580-2

*Putnam, A. L., & Roediger, H. L. III (2013). Does response mode affect amount recalled or the magnitude of the testing effect? *Memory & Cognition*, *41*, 36–48. doi:10.3758/s13421-012-0245-x

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

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

*Pyc, M. A., & Rawson, K. A. (2011). Costs and benefits of dropout schedules of test-restudy practice: Implications for student learning. *Applied Cognitive Psychology*, *25*, 87–95. doi:10.1037/a0026166

*Pyc, M. A., & Rawson, K. A. (2012). Why is test-restudy practice beneficial for memory? An evaluation of the mediator shift hypothesis. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *38*, 737–746. doi:10.1037/a0026166

*Pyke, A. A., & LeFevre, J. (2011). Calculator use need not undermine direct-access ability: the roles of retrieval, calculation, and calculator use in the acquisition of arithmetic facts. *Journal of Educational Psychology*, *103*, 607–616. doi:10.1037/a0023291

Raaijmakers, J. G. W., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review*, *88*, 93–134. doi:10.1037/0033–295X.88.2.93

*Richland, L. E., Kornell, N., & Kao, L. S. (2009). The pretesting effect: Do unsuccessful retrieval attempts enhance learning? *Journal of Experimental Psychology: Applied*, *15*, 243–257. doi:10.1037/a0016496

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

Roediger, H. L., III, & Butler, A. C. (2013). Retrieval practice (testing) effect. In H. L. Pashler (Ed.), *Encyclopedia of the mind* (pp. 660–661). Thousand Oaks, CA: Sage.

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

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

*Roediger, H. L., III, & Marsh, E. J. (2005). The positive and negative consequence of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 1155–1159. doi:10.1037/0278-7393.31.5.1155

Roediger, H. L., III, & Pyc, M. A. (2012). Applying cognitive psychology to education: Complexities and prospects. *Journal of Applied Research in Memory and Cognition*, *1*, 242–248. doi:10.1016/j.jarmac.2012.09.002

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

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

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin*, *86*, 638–641. doi:10.1037/0033–2909.86.3.638

Rothman, R. (2011). *Something in common: The Common Core Standards and the next chapter in American education*. Cambridge, MA: Harvard Education Press.

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (Eds.). (2005). *Publication bias in meta–analysis: Prevention, assessment, and adjustment*. New York, NY: Wiley.

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. doi:10.1037/a0037559

*Royer, J. M. (1973) Memory effects for test-like events during acquisition of foreign language vocabulary. *Psychological Reports*, *32*, 195–198. doi:10.2466/pr0.1973.32.1.195

*Saville, B. K., Pope, D., Lovaas, P., & Williams, J. (2012) Interteaching and the testing effect: A systematic replication. *Teaching of Psychology*, *39*, 280–283. doi:10.1177/0098628312456628

Shea, J. B., & Morgan, R. L. (1979). Contextual interference effects on the acquisition, retention, and transfer of a motor skill. *Journal of Experimental Psychology: Human Learning and Memory*, *5*, 179–187. doi:10.1037/0278–7393.5.2.179

*Smith, T. A., & Kimball, D. R. (2010). Learning from feedback: Spacing and the delay–retention effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*, 80–95. doi:10.1037/a0017407

*Smith, M. A., Roediger, H. L., III, & Karpicke, J. D. (2013). Covert retrieval practice benefits retention as much as overt retrieval practice. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *39*, 1712–1725. doi:10.1037/a0033569

Spitzer, B., & Bäuml, K.-H. (2009). Retrieval-induced forgetting in a category recognition task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 286–291 10.1037/a0014363

*Sumowski, J. F., Chiaravalloti, N., & DeLuca, J. (2010). Retrieval practice improves memory in multiple sclerosis: Clinical application of the testing effect. *Neuropsychology*, *24*, 267–272. doi:10.1037/a0017533

*Sumowski, J. F., Wood, H. G., Chiaravalloti, N., Wylie, G. R., Lengenfelder, J., & DeLuca, J. (2010). Retrieval practice: A simple strategy for improving memory after traumatic brain injury. *Journal of the International Neuropsychological Society*, *16*, 1147–1150. doi:10.1017/S1355617710001128

*Szpunar, K. K., McDermott, K. D., & Roediger, H. L. (2007). Expectation of a final cumulative test enhances long-term retention. *Memory & Cognition*, *35*, 1007–1013. doi:10.3758/BF03193473

*Szpunar, K. K., McDermott, K. B., & Roediger, H. L. (2008). Testing during study insulates against the build-up of proactive interference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 1392–1399. doi:10.1037/a0013082

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Boston, MA: Pearson.

Tobin, K. (1987). The role of wait time in higher cognitive level learning. *Review of Educational Research*, *57*, 69–95. doi:10.3102/00346543057001069

*Toppino, T. C., & Cohen, M. S. (2009). The testing effect and the retention interval. *Experimental Psychology*, *56*, 252–257. doi:10.1027/1618-3169.56.4.252

*Tse, C., Balota, D. A., & Roediger, H. L. (2010). The benefits and costs of repeated testing on the learning of face–name pairs in healthy older adults. *Psychology and Aging*, *25*, 833–845. doi:10.1037/a0019933

*Tuckman, B. W., & Trimble, S. (1997, August). *Using tests as a performance incentive to motivate eighth-graders to study*. Paper presented at the 105th Annual Meeting of the American Psychological Association, Chicago, IL.

*van Gog, T., & Kester, L. (2012). A test of the testing effect: Acquiring problem-solving skills from worked examples. *Cognitive Science*, *36*, 1532–1541. doi:10.1111/cogs.12002

*Verkoeijen, P. P. J. L., Bouwmeester, S., & Camp, G. (2012). A short-term testing effect in cross-language recognition. *Psychological Science*, *23*, 567–571. doi:0.1177/0956797611435132

*Verkoeijen, P. P. J. L., Tabbers, H. K., & Verhage, M. L. (2011). Comparing the effects of testing and restudying on recollection in recognition memory. *Experimental Psychology*, *58*, 490–498. doi:10.1027/1618-3169/a000117

Vinovskis, M. (2008). *From a Nation at Risk to No Child Left Behind: National education goals and the creation of federal education policy*. New York, NY: Teachers College Press.

*Vojdanoska, M., Cranney, J., & Newell, B. R. (2009). The testing effect: the role of feedback and collaboration in a tertiary classroom setting. *Applied Cognitive Psychology*, *24*, 1183–1195. doi:10.1002/acp.1630

*Weinstein, Y., McDermott, K. B., & Roediger, H. L. (2010). A comparison of study strategies for passages: Rereading, answering questions, and generating questions. *Journal of Experimental Psychology: Applied*, *16*, 308–316. doi:10.1037/a0020992

*Wenger, S. K., Thompson, C. P., & Bartling, C. A. (1980). Recall facilitates subsequent recognition. *Journal of Experimental Psychology: Human Learning and Memory*, *6*, 135–144.

*Zaromb, F. M., & Roediger, H. L. (2010). The testing effect in free recall is associated with enhanced organizational processes. *Memory & Cognition*, *38*, 995–1008. doi:10.3758/MC.38.8.995

Authors

OLUSOLA O. ADESOPE, PhD, is an associate professor and coordinator of the educational psychology program in the Department of Educational Leadership, Sports Studies, and Educational/Counseling Psychology, Washington State University, Pullman, WA 99164, USA; email: *olusola.adesope@wsu.edu*. His research is at the intersection of educational psychology, learning sciences, and instructional design and technology. His recent research focuses on the cognitive and pedagogical underpinnings of learning with computer-based multimedia resources, knowledge representation through interactive concept maps, meta-analysis of empirical research, and advancing learning, instructional principles and assessments in science, technology, engineering, and mathematics.

DOMINIC A. TREVISAN is a doctoral student in Educational Psychology at Simon Fraser University, 8888 University Drive, Burnaby, British Columbia, Canada V5A 1S6; email: *dtrevisa@sfu.ca*. Some of his research interests include applications of cognitive psychology in the classroom, social and emotional learning, and autism.

NARAYANKRIPA SUNDARARAJAN is a graduate research assistant in the Department of Educational Leadership, Sports Studies, and Educational/Counseling Psychology, Washington State University, Cleveland 80, Pullman, WA 99164, USA; email: *n.sundararajan@wsu.edu*. Her research interests include educational psychology, instructional techniques, multimedia, and research methods.