

Testing (Quizzing) Boosts Classroom Learning: A Systematic and Meta-Analytic Review

Chunliang Yang and Liang Luo
Beijing Normal University

Miguel A. Vadillo
Universidad Autónoma de Madrid

Rongjun Yu
National University of Singapore

David R. Shanks
University College London

Over the last century hundreds of studies have demonstrated that testing is an effective intervention to enhance long-term retention of studied knowledge and facilitate mastery of new information, compared with restudying and many other learning strategies (e.g., concept mapping), a phenomenon termed *the testing effect*. How robust is this effect in applied settings beyond the laboratory? The current review integrated 48,478 students' data, extracted from 222 independent studies, to investigate the magnitude, boundary conditions, and psychological underpinnings of test-enhanced learning in the classroom. The results show that overall testing (quizzing) raises student academic achievement to a medium extent ($g = 0.499$). The magnitude of the effect is modulated by a variety of factors, including learning strategy in the control condition, test format consistency, material matching, provision of corrective feedback, number of test repetitions, test administration location and timepoint, treatment duration, and experimental design. The documented findings support 3 theories to account for the classroom testing effect: additional exposure, transfer-appropriate processing, and motivation. In addition to their implications for theory development, these results have practical significance for enhancing teaching practice and guiding education policy and highlight important directions for future research.




Public Significance Statement

Testing (class quizzing) yields a variety of learning benefits, even though learners, instructors, and policymakers tend to lack full metacognitive insight into the virtues of testing. The current meta-analysis finds a reliable advantage of testing over other strategies in facilitating learning of factual knowledge, concept comprehension, and knowledge application in the classroom. Overall, testing is not only an assessment *of* learning but also an assessment *for* learning.

Keywords: academic achievement, meta-analysis, motivation, testing effect, transfer-appropriate processing

Learning is defined as mastering new information or skills or modifying existing knowledge. The question of how to optimize learning has long been a fundamental concern for learners, instructors, policymakers, and researchers. Many effective strategies have been identified, such as note-taking, using mnemonics, constructivist concept maps, rereading, interleaving to-be-learned materials,

and so on (for a review, see Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013). Although testing is usually administered for assessment purposes (e.g., measuring comprehension or teaching effectiveness) in educational settings, research has repeatedly demonstrated that retrieval practice (i.e., retrieving information from memory) can more effectively consolidate long-term retention of

Chunliang Yang, Institute of Developmental Psychology, Faculty of Psychology, Beijing Normal University;  Liang Luo, Institute of Developmental Psychology, Faculty of Psychology and Collaborative Innovation Centre of Assessment Toward Basic Education Quality, Beijing Normal University;  Miguel A. Vadillo, Departamento de Psicología Básica, Universidad Autónoma de Madrid; Rongjun Yu, Department of Psychology, National University of Singapore;  David R. Shanks, Division of Psychology and Language Sciences, University College London.

The data and analysis code are publicly available at <https://osf.io/ewrsd/>.

This research was supported by the Fundamental Research Funds for the Central Universities (2019NTSS28), the Natural Science Foundation of China (31671130; 32000742), the NUS FASS RSB-PDF Scheme under WBS (C-581-000-207-532; C-581-000-777-532), the United Kingdom Economic and Social Research Council (ES/S014616/1), and the MadrI+D Science Foundation (2016-T1/SOC-1395).

Correspondence concerning this article should be addressed to Rongjun Yu, Department of Psychology, National University of Singapore, Arts and Social Science Building 4, 1 Lower Kent Ridge Road, Singapore 117570, Singapore. E-mail: psyjr@nus.edu.sg

studied information (Roediger & Karpicke, 2006a) and facilitate mastery of new information (Yang, Potts, & Shanks, 2018) by comparison with other strategies (e.g., concept mapping: Karpicke & Blunt, 2011; restudying: Roediger & Karpicke, 2006b; note-taking: Rummel, Schweppe, Gerst, & Wagner, 2017). This phenomenon is termed the *testing effect* or *test-enhanced learning*.

Test-enhanced learning has received considerable research attention over the last century (Abbott, 1909), and the past decade has witnessed an exponential increase in explorations of its effectiveness and boundary conditions (see Figure 1 for an illustration of the substantial increase in publications across the past 50 years). Notably, many studies have investigated the applicability of test-enhanced learning in the classroom (e.g., Chan, Kim, Garavalia, & Wang, 2018; Gokcora & DePaulo, 2018; Poljičanin et al., 2009). In these classroom studies, students in the intervention condition took regular quizzes by recalling or applying studied information to solve new problems, and quizzes were administered in a variety of formats, such as multiple choice, fill-in-the-blank, cued recall, short answer, free recall, short essay, and so on.¹ By contrast, students in the control condition were not quizzed. In course exams administered at the end of the semester or academic year, students in the intervention condition typically outperformed those in the control condition (for a review, see Moreira, Pinto, Starling, & Jaeger, 2019).

The current article aims to (a) provide a brief literature review of the testing effect, (b) offer a comprehensive meta-analytic review of its applications in the classroom, (c) explore the mechanisms underlying the classroom testing effect, and (d) illuminate some directions for future research.

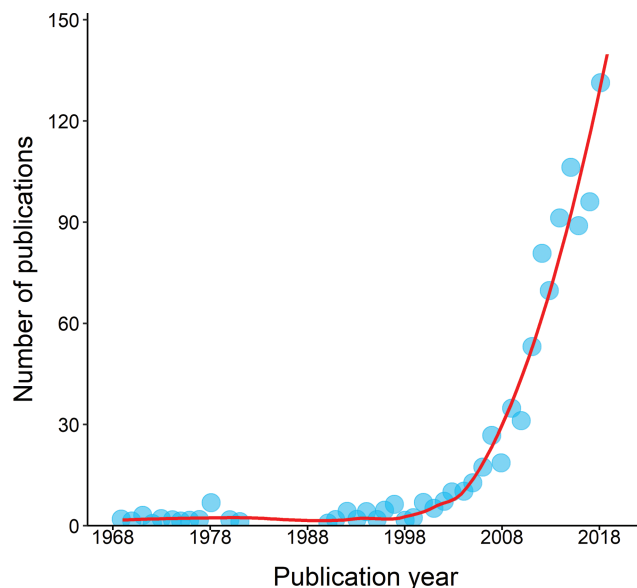


Figure 1. Number of publications exploring test-enhanced learning from 1968 to 2018. Data were extracted from a Web of Science search conducted on May 7th 2019, using the term ["testing effect" OR "test-enhanced learning" OR "test-potentiated learning" OR "retrieval practice"]. The red line represents the developing trend of publications across years. See the online article for the color version of this figure.

Test-Enhanced Learning and Metacognitive Insight

Test-Enhanced Consolidation of Studied Information (the Backward Testing Effect)

In educational settings, testing is usually regarded as an evaluative instrument to assess learning and comprehension, or to gauge learners' ongoing progress toward learning objectives. However, a large body of research has supplied convincing evidence that testing is also an effective technique to facilitate long-term retention of studied information (for reviews, see Adesope, Trevisan, & Sundararajan, 2017; Roediger & Karpicke, 2006a; Roediger, Putnam, & Smith, 2011; Rowland, 2014), a phenomenon we term the *backward testing effect* to make it distinct from the *forward testing effect* (see below for details; Pastötter & Bäuml, 2014; Yang et al., 2018). The backward testing effect (i.e., testing consolidates retention of studied information) is a robust phenomenon across different educational materials (such as foreign-translation word pairs, text passages, and lecture videos) in both the laboratory and the real classroom.

Two examples are illustrative. Roediger and Karpicke (2006b) asked participants to study two text passages, with one passage studied twice and the other studied once and tested once. In a final test one week later, the tested passage was substantially better recalled than the restudied one, demonstrating the backward testing effect. Leeming (2002b) documented test-enhanced learning in the classroom. In a 5-week Learning and Memory course, an exam-a-day group took a short quiz after each day's classes, in which students answered two short-answer questions and were provided with 2–3 min of corrective feedback. The exam-a-day group took about 20 exams in total across the whole summer term. By contrast, a three-exam group only took exams on three classes across the term. In a final course exam administered near the end of the term, the exam-a-day group achieved significantly higher grades than the three-exam group.

Test-Enhanced Learning of New Information (the Forward Testing Effect)

The above discussion mainly focuses on the phenomenon that testing consolidates long-term retention of studied information. Besides aiding the mastery of studied information, many experiments have documented that testing on studied information can also facilitate acquisition of new information, a phenomenon termed the forward testing effect or test-potentiated new learning (i.e., testing prospectively facilitates learning of new information; for reviews, see Chan, Meissner, & Davis, 2018; Pastötter & Bäuml, 2014; Yang et al., 2018). For instance, Szpunar, McDermott, and Roediger (2008) instructed two groups of participants to study five 18-word lists. The experimental group took a free recall test shortly after studying each list, while a control group solved math problems (a filler task) after studying each of Lists 1–4 but took a test on List 5. In the free recall test on List 5, the experi-

¹ These quizzes are typically designed to elicit recall or apply studied information to solve new problems. Tests requiring subjective reports (e.g., attitudes toward curriculum content) as well as intelligence and other standardized (e.g., creativity) tests are not the focus of the current review and hence are not discussed further.

mental group correctly recalled twice as many words as the control group, indicating a strong forward testing effect (i.e., interim testing on Lists 1–4 doubled learning and recall of List 5).

Although this forward effect has only been identified recently, a wealth of research has explored its generalizability and limits. It has been established that the effect is robust across a variety of educational materials, such as word lists (Aslan & Bäuml, 2016; Bäuml & Kliegl, 2013; Nunes & Weinstein, 2012; Pastötter, Schicker, Niedernhuber, & Bäuml, 2011; Pierce, Gallo, & McCain, 2017; Weinstein, Gilmore, Szpunar, & McDermott, 2014; Yang, Potts, & Shanks, 2017), line drawings of common objects (Pastötter, Weber, & Bäuml, 2013), foreign-translation word pairs (Cho, Neely, Crocco, & Vitrano, 2017; Yang et al., 2017), face-name pairs (Weinstein, McDermott, & Szpunar, 2011; Yang et al., 2017), text passages (Healy, Jones, Lalchandani, & Tack, 2017; Wissman, Rawson, & Pyc, 2011; Zhou, Yang, Cheng, Ma, & Zhao, 2015), lecture videos (Jing, Szpunar, & Schacter, 2016; Szpunar, Khan, & Schacter, 2013; Yue, Soderstrom, & Bjork, 2015), artistic styles (Lee & Ahn, 2018; Yang & Shanks, 2018), and spatial episodic information (Bufe & Aslan, 2018). The effect is not limited to healthy young adults but also occurs in children (Aslan & Bäuml, 2016), older adults (Pastötter & Bäuml, 2019), and patients with traumatic brain injury (Pastötter et al., 2013). Moreover, it generalizes to individuals with different levels of working memory capacity and test anxiety (Yang et al., 2020).

Insight and Application of Test-Enhanced Learning

Although the beneficial effects of testing on studied and new information are broad, researchers have frequently expressed dismay that learners, instructors, and policymakers tend not to appreciate these benefits and that retrieval practice has not been applied to enhance educational practices as widely as it could be (Roediger & Karpicke, 2006b). In a review article on effective learning strategies, Dunlosky et al. (2013, p. 29) wrote that “[W]e suspect that most students would prefer to take as few tests as possible.”

Roediger and Karpicke (2006b) observed that although tested passages were better recalled than restudied ones in their experiment on a test one week later, participants judged that the restudied passages would be better remembered than the tested ones. More recent findings come from Kirk-Johnson, Galla, and Fraundorf (2019, Experiment 3). In this study, participants studied two passages, with one restudied and the other tested. Next they studied a third passage, decided whether they would like to restudy or take a test on it, and employed the selected strategy to reprocess it. Finally, 48 h later they took a cumulative test on all three passages. The results showed that a majority of participants chose to restudy the third passage, even though those who chose to take a practice test on it recalled it better in the cumulative test than those who chose to restudy it.

Some questionnaire surveys document similar findings. In a survey conducted by Karpicke, Butler, and Roediger (2009), only 1% of participants (students from the University of Washington at St. Louis) regarded retrieval practice as their best study strategy, and only 11% reported that they self-administered tests while studying. It is worth noting that some surveys found that students do employ self-testing during self-regulated learning (Geller et al., 2018; Hartwig & Dunlosky, 2012; Kornell & Bjork, 2007; McAndrew, Morrow, Atiyeh, & Pierre, 2016; Yan, Thai, & Bjork, 2014).

In all of the above-cited studies, which used the same questionnaire, students were prompted to answer “If you quiz yourself while you study, why do you do so?” by selecting one from four options: *A. I learned more from that way than rereading; B. To figure out how well I have learned the information I’m studying; C. I found quizzing is more enjoyable than rereading; D. I do not quiz myself.* These studies consistently demonstrated that only a minority of students (e.g., 9% in Kornell & Bjork, 2007) selected Option D. However, these studies also consistently showed that the majority of students (e.g., 68% in Kornell & Bjork, 2007) administered self-tests to determine how well they had mastered studied information (B), whereas only a minority (18%) acknowledged that testing facilitates learning (A), implying that self-tests are taken for diagnostic purposes rather than driven by metacognitive appreciation of test-enhanced learning (for related discussion, see Agarwal, D’Antonio, Roediger, McDermott, & McDaniel, 2014).

Similar to students, some teachers may also lack full appreciation of the benefits of test-enhanced learning. For instance, Morehead, Rhodes, and DeLozier (2016) prompted 76 college teachers to state why they thought students should test themselves. The majority (68%) of them reported that “students should administer tests to figure out how well they have learned the information they are studying,” with only 19% reporting that “students will learn more through testing than rereading.” Besides metacognitive unawareness, some instructors are reluctant to incorporate quizzes and exams into curricula, because they believe (not unreasonably) that administering quizzes is time-consuming (Roediger & Karpicke, 2006a) and that scoring is excessively demanding. Some policymakers also underappreciate the benefits of testing or erroneously (from the perspective of student attainment) value other teaching activities over testing and advocate minimizing its use in schools. For instance, the Singapore Ministry of Education recently proposed to reduce the number of assessments at primary and secondary schools to free up more time for other goals (Singapore Ministry of Education, 2018). Other factors that may potentially result in underemployment of test-enhanced learning in educational settings are considered in the Discussion.

In summary, testing can enhance learning of studied and new information. Nonetheless, students, teachers, and policymakers tend to lack metacognitive insight into the virtues of testing or may inappropriately value other activities more than testing, leading to its underemployment in educational settings. If a comprehensive review of test-enhanced learning confirms its effectiveness in the classroom, this may serve to raise practitioners’ appreciation and promote application of testing in educational settings.

Underlying Mechanisms of Test-Enhanced Learning

Dozens of explanations have been proposed to account for test-enhanced learning. For instance, Rowland (2014) and Karpicke (2017) summarized a variety of explanations of how testing can enhance long-term retention of studied information, and Yang et al. (2018) reviewed eight theories proposed to account for why testing of studied information potentiates learning of new information. For the sake of brevity, here we combine several of the most influential explanations into four major accounts. At the outset, it should be acknowledged that these explanations are not necessarily mutually exclusive, and the mechanisms proposed by

some of them may combine to produce overlapping learning enhancements in some situations (Yang et al., 2018).

Additional Exposure

In many early studies, the testing effect was explored by comparing testing with a filler task or no activity (e.g., Glover, 1989). In these studies, the learned materials in the treatment condition were tested following initial learning, but testing was replaced by a filler (distractor) task or there were no further activities in the control condition. Because the learned content was not reviewed in the control condition, the testing effect documented in these studies could be explained by an *additional exposure* theory, which simply assumes that the benefits of retrieval derive from the reexposure duration (study time) it provides to the successfully recalled materials (Slamecka & Katsaiti, 1988; Thompson, Wenger, & Bartling, 1978).

This view has been challenged, however, by numerous later findings showing that testing can also more effectively boost long-term retention even when compared with restudying (which matches exposure time; e.g., Adesope et al., 2017; Roediger & Karpicke, 2006b; Rowland, 2014). Regardless of this challenge, the additional exposure theory remains a viable account for many important findings. For instance, the magnitude of test-enhanced learning is substantially larger when comparing testing with a filler task or no activity (Hedges' $g = 0.93$) than when comparing it with restudying ($g = 0.51$; Adesope et al., 2017), and testing followed by feedback more effectively enhances later recall ($g = 0.73$) than it does without feedback ($g = 0.39$; Rowland, 2014). Overall, the above discussion suggests that additional exposure is likely to be one but not the only explanation for test-enhanced learning.

Retrieval Effort

Retrieval effort theory proposes that test-enhanced learning arises from cognitive effort expended in retrieving studied information from memory (Glover, 1989; Karpicke & Roediger, 2007a; Pyc & Rawson, 2009). This theory follows R. A. Bjork's "desirable difficulty" framework (Bjork, 1994; Bjork & Bjork, 2011), which assumes that a difficult/effortful learning process produces better retention than an easy or less effortful one. R. A. Bjork proposed that memories have two strengths: storage strength, defined as the long-term establishment of memory, and retrieval strength, representing a memory's momentary accessibility. For an item with low retrieval accessibility/strength (e.g., an item retrieved after a long retention interval), its successful retrieval from memory is demanding and effortful, which in turn boosts its long-term storage strength. Accordingly, a prediction of the retrieval effort theory is that, by comparison with easy retrieval practice, difficult practice will be more beneficial for long-term retention (Karpicke & Roediger, 2007a; Pyc & Rawson, 2009; Stenlund, Sundström, & Jonsson, 2016; Wang & Zhao, 2019).

Supporting evidence for the retrieval effort theory comes from Rowland's (2014) meta-analysis, which found that recall tests, associated with greater retrieval difficulty, produced larger memory gains ($g = 0.72$ for cued recall and $g = 0.82$ for free recall) than recognition tests ($g = 0.36$ for multiple choice and old/new recognition). It is, however, worth noting that Adesope et al.'s (2017) meta-analysis found that multiple-choice tests ($g = 0.70$)

produced greater mnemonic benefits than cued recall ($g = 0.58$), free recall ($g = 0.62$), and short answer tests ($g = 0.48$), inconsistent with the retrieval effort view.

Transfer-Appropriate Processing

The retrieval effort explanation focuses on the direct benefit of testing—the idea that retrieval practice directly consolidates memories. Besides this, testing also confers a variety of indirect (mediating) advantages. For example, prior testing may inform learners about the test format and teach them how to "learn to the test." Accordingly, they may adjust their subsequent encoding and retrieval strategies (Chan, Manley, Davis, & Szpunar, 2018; Thomas & McDaniel, 2013; Yang & Shanks, 2018), promoting efficient learning and benefiting retrieval.

Another potential indirect benefit of testing is *transfer-appropriate processing*, a concept referring to the well-established phenomenon whereby recall performance depends on the similarity between the mental operations brought to bear during the acquisition and assessment phases (Blaxton, 1989; Morris, Bransford, & Franks, 1977). According to the transfer-appropriate processing theory, the reason why retrieval practice is beneficial for later retrieval is that initial tests and final assessments share similar mental processes, namely the processing steps required to retrieve a specific piece of information from memory. A major prediction of this theory is that retrieval practice ought to produce larger learning gains when the test format in the acquisition phase and final assessment are matched than when they are mismatched. It is worth noting that this prediction has been placed under doubt by recent research (e.g., Karpicke, Lehman, & Aue, 2014; Rowland, 2014). For instance, Rowland's (2014) meta-analysis documented that test format consistency did not significantly modulate the testing effect. Moreover, Carpenter and Delosh (2006) observed that cued-recall tests in the acquisition phase generated larger learning gains than recognition tests, regardless of whether the test format in the final assessment phase was cued recall or recognition (for related findings, see Glover, 1989; Kang, McDermott, & Roediger, 2007).

Despite these challenges, there is also a variety of supporting evidence. For instance, in contrast to the studies discussed above, Veltre, Cho, and Neely (2015) did find support for the predicted interaction between test format at acquisition and final test, and Adesope et al.'s (2017) meta-analysis found a larger testing effect for matched than for mismatched formats (for related findings, see Duchastel & Nungester, 1982). In brief, two meta-analyses yielded inconsistent findings regarding this theory (Adesope et al., 2017; Rowland, 2014), and further theoretical exploration is required.

Motivation

The class of *motivation* theories, which claim that frequent tests motivate learners to sustain or enhance their efforts to learn, comprises another influential explanation. Testing can maintain/enhance learning motivation in several different ways (Yang et al., 2018). For instance, experiencing retrieval failures in prior tests may induce metacognitive awareness of the difficulty of achieving successful retrieval and lead to dissatisfaction about poor test performance, which in turn may drive learners to study harder (Cho et al., 2017; Yang et al., 2018). Learners are frequently

overconfident regarding their learning status, and test performance provides diagnostic feedback to inform them about the gap between their anticipated and actual learning level (Szpunar, Jing, & Schacter, 2014), which then motivates them to expend more effort to narrow the perceived gap. In addition, frequent tests may induce high test expectancy (i.e., expecting to be tested subsequently), and test expectancy is an important motivator driving students to commit more effort to prepare for subsequent tests (Agarwal & Roediger, 2011; Szpunar, McDermott, & Roediger, 2007; Yang, Chew, Sun, & Shanks, 2019).

Numerous studies have provided supporting evidence for the motivation explanation. For instance, Schrank (2016) found that class quizzes increase attendance; Heiner, Banet, and Wieman (2014) reported that a preannounced quiz encourages students to read the assigned textbook material and prepare better before class; Yang et al. (2017) showed that frequent tests drive learners to allocate more time to learning; Szpunar et al. (2013) observed that learners make more notes when they are frequently tested; Jing et al. (2016) found that frequent tests reduce task-unrelated thoughts (i.e., mind wandering) while watching lecture videos; and Weinstein et al. (2014) found that frequent tests induce high test expectancy which in turn boosts test performance. However, it is noteworthy that Kang and Pashler (2014) reported that motivational interventions (e.g., monetary incentives) fail to significantly alter the magnitude of the testing effect, implying minimal influence of learning motivation on test-enhanced learning. A large-scale meta-analysis is required to clarify the role of motivation in the classroom testing effect. The relationship between the role of motivation in test-enhanced learning and the wider and very influential literature on motivation in the classroom has been underexplored. We return to this issue in the Discussion.

Rationale of the Current Meta-Analytic Review

To our knowledge, at least 90 review articles, identified in our literature search (see below), have been published endeavoring to summarize the findings of research on the testing effect (e.g., Moreira et al., 2019; Roediger & Karpicke, 2006a; Roediger, Putnam, et al., 2011; Yang et al., 2018). Nonetheless, a majority (92%; 83 of 90) of them were qualitative literature reviews, and only seven conducted and reported formal meta-analyses (Adesope et al., 2017; Bangert-Drowns, Kulik, & Kulik, 1991; Chan, Manley, et al., 2018; Pan & Rickard, 2018; Phelps, 2012; Rowland, 2014; Schwieren, Barenberg, & Dutke, 2017). Five of these mainly integrated research findings from laboratory studies, with far fewer from classroom research (Adesope et al., 2017; Chan, Manley, et al., 2018; Pan & Rickard, 2018; Phelps, 2012; Rowland, 2014). For instance, Rowland (2014) explicitly excluded classroom studies from his meta-analysis, while only 11% of the 272 effects included in Adesope et al.'s (2017) recent meta-analysis came from classroom research.

It is important to highlight that there are numerous examples demonstrating that laboratory research findings do not always generalize to the classroom (e.g., Lundeberg & Fox, 1991) because laboratory settings are different from those in the classroom in many important respects. For instance, laboratory research is typically implemented on personal computers and is unimodal (e.g., words studied visually). In contrast, courses are orally taught by teachers, involve multimodal learning, and class quizzes and

course exams are typically hand-written. In the laboratory, researchers strictly control irrelevant variables to specifically investigate the impacts of testing (such as presenting all materials in a random order at a fixed pace and counterbalancing material assignment to avoid item selection effects). By contrast, the classroom environment is more complex and noisier (often literally), and many variables strictly controlled in laboratory research are unlikely to be controlled in the classroom.

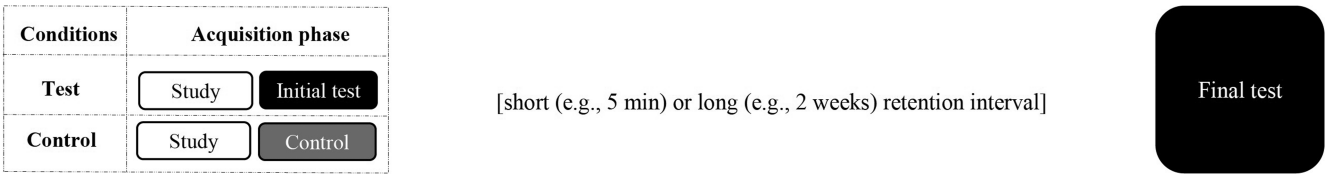
Most laboratory studies (81% of the effects in Rowland's meta-analysis) used word pairs and word lists as their principal stimuli, which are somewhat unrepresentative of educational materials (such as textbooks and lecture content). For laboratory research, the most widely used test formats are cued recall and free recall (88% in Rowland's meta-analysis), but multiple-choice tests are the cornerstone of assessment in educational settings (e.g., Bjork, Little, & Storm, 2014; Roediger, Agarwal, McDaniel, & McDermott, 2011). About 67% of the effects in Rowland's meta-analysis did not offer corrective feedback, but it is unusual in the classroom to administer a formative quiz without providing feedback. The amount of information for students to master is substantially greater in the classroom than in a single laboratory experiment (Roediger & Karpicke, 2006a). It is also important to highlight that students are typically more motivated to study in the classroom than participants in the laboratory (Kang & Pashler, 2014). Students frequently have background knowledge about class materials, but participants commonly have little prior knowledge about learning materials (e.g., Swahili-English word pairs) employed in laboratory research (Carpenter et al., 2016).

Most importantly, there are extensive divergences in the research procedures between laboratory and classroom settings (as schematically illustrated in Figures 2A and 2B, respectively). In laboratory research, participants typically study some materials and then take an initial test or complete a control activity. After either a short (e.g., 5 min) or long (e.g., one month) retention interval, a final test is administered (for an example, see Roediger & Karpicke, 2006b). In classroom research, by contrast, students typically take a quiz following each class (or unit) across a whole semester, in which they repeatedly experience study-test cycles with different materials studied in each cycle. They finally take a semester exam to measure their attainment (for an example, see Roediger, Agarwal, et al., 2011). These procedural divergences are critical and may well mean that different mechanisms underlie testing effects in the laboratory and classroom.

As shown in Figure 2B, the classroom procedure is similar to the widely used multiblock procedure for exploring the forward testing effect (see Yang et al., 2018, Figure 1; also see the above discussion of Szpunar et al., 2008). Hence, the classroom testing effect may originate from both direct (i.e., testing consolidates studied information) and indirect forward (i.e., testing boosts new learning) benefits of testing, whereas the laboratory testing effect, which typically involves only a single study-test cycle (see Figure 2A), may largely result from the direct benefit.

All of these divergences between laboratory and classroom research may significantly modulate the testing effect. Take corrective feedback as an illustration. Many studies have established that corrective feedback is a fortifier of test-enhanced learning that increases the benefits of retrieval (Kang et al., 2007; Lantz & Stawiski, 2014; Moreira et al., 2019; Rowland, 2014; Vojdanoska, Cranney, & Newell, 2010). Given that corrective feedback is frequently not provided in

A: Laboratory research design



B: Classroom research design

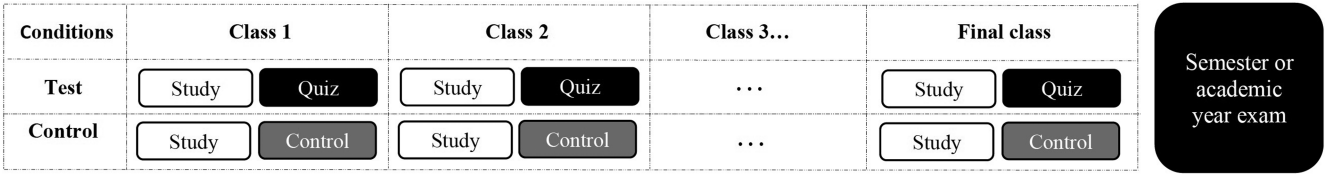


Figure 2. (A) Widely used design in laboratory research, in which participants study some materials (e.g., word pairs), take or do not take initial test(s), and then complete a final test to measure retention. (B) Common design in classroom research. Students take or do not take a quiz after a class (or unit) and this study-test cycle repeats until the end of the semester (or academic year). (Note that different materials are studied in each cycle.) Finally, they complete a semester or academic year exam to measure attainment.

laboratory research (e.g., 67% of laboratory studies in Rowland’s meta-analysis) but typically offered in the classroom, the magnitude of the classroom testing effect may be greater than that seen in the laboratory. However, there are also reasons to expect a smaller or even null testing effect in the classroom. For instance, based on the motivation theory, testing boosts learning through enhancing motivation. Because students in the classroom are typically more motivated to study than participants in the laboratory (Kang & Pashler, 2014), there may be little room left for testing to improve learning motivation in the classroom, leading to a smaller or null testing benefit.

In sum, laboratory and classroom research differ in many important ways, and laboratory research often lacks ecological validity. Hence, it is unknown to what extent laboratory findings can inform educational practices and pedagogical policies, and a comprehensive meta-analysis especially focusing on classroom research is called for. Previously, two meta-analytic reviews have been conducted to specifically focus on the effect of testing on classroom learning: Bangert-Drowns et al. (1991) and Schwieren et al. (2017). However, Bangert-Drowns et al.’s meta-analysis was published nearly 30 years ago and focused on the test frequency effect (that is, more frequently administered quizzes produce greater learning enhancement than less frequent ones), whereas Schwieren et al.’s (2017) meta-analytic review only focused on the testing effect in Psychology classes. Because of the restriction to the subject of Psychology, the majority of the learners included in Schwieren et al.’s meta-analysis were college students, with a minority coming from high schools. Because test-enhanced learning may be study material and age dependent, it remains unclear whether the classroom testing effect can be extended to other subjects (such as Biology, Engineering, and so on) and to younger age groups (such as elementary school children). In addition, as we will show below, many important moderators of the classroom testing effect were not investigated by Bangert-Drowns et al. (1991) or Schwieren et al. (2017).

Another important reason why a new meta-analysis may be valuable is that many results emerging from prior meta-analyses were inconsistent or even conflicting. For instance, Rowland’s (2014) meta-analysis showed that offering corrective feedback following testing almost doubled the testing benefit ($g = 0.73$ with feedback compared with $g = 0.39$ without feedback), but Adesope et al. (2017) found no moderating effect of feedback (with feedback, $g = 0.63$; without feedback, $g = 0.60$). Such inconsistent findings are problematic theoretically as well as confusing for practitioners regarding whether corrective feedback should be presented or withheld following class quizzing. Rowland (2014) and Adesope et al. (2017) also reported inconsistent results regarding whether test format consistency modulates test-enhanced learning, which bears theoretical importance for the transfer-appropriate processing account. Hence, further explorations of these important questions are undertaken in the current review. Going beyond Rowland (2014) and Adesope et al. (2017), the current review directly investigated these questions in educational settings (i.e., the classroom) with a substantially larger dataset (i.e., more than 48,000 students’ data extracted from more than 200 studies).

Lastly, but importantly, exploring the moderators of the classroom testing effect provides a pathway to explore its underlying mechanisms. It is possible that different mechanisms may contribute to testing effects in the laboratory and classroom (see above for a detailed discussion). Previous theoretical explorations, derived mainly from laboratory research, might have limited applicability to the classroom testing effect. Unfortunately, most classroom studies have concentrated on the application of test-enhanced learning in the classroom, with its psychological underpinnings largely unexamined. Without a much deeper exploration of its cognitive and motivational underpinnings and boundary conditions in the classroom, educational translation and exploitation are likely to be hindered.

Research Questions

The current review aims to address 19 important questions about the classroom testing effect. As an aid to the reader, we summarize these questions in Table 1 (left side) with the corresponding answers (meta-analytic findings; right side). In addition, the Results section is organized in the same order as they appear in the table.

Q.1 Does (and If So, to What Extent) Classroom Testing Boost Student Attainment?

The primary aim of the current review is to integrate existing research data to explore whether (and if so, to what extent) class quizzes potentiate student attainment. To foreshadow, the

Table 1

Questions Explored in the Current Review and the Corresponding Research Findings

Question	Answers (research findings)
Q1. Does (and if so, to what extent) classroom testing boost student attainment?	Classroom testing significantly boosts student learning achievement to a medium extent ($g = 0.499$). The p -curve analysis shows strong evidence supporting the existence of test-enhanced learning in the classroom. The risk of publication bias in this research field is small.
Q2. Against what comparison treatments does quizzing enhance learning?	By comparison with no/filler activity ($g = 0.610$), testing with fewer questions ($g = 0.465$), restudying ($g = 0.330$), and other elaborative strategies ($g = 0.095$), testing is overall a more powerful method to enhance classroom learning.
Q3. Does quiz format matter?	Test-enhanced learning generalizes to different test formats ($g = 0.913$ for Matching; $g = 0.773$ for Fill-in-the-blank; $g = 0.638$ for Short answer; $g = 0.567$ for Multiple choice; $g = 0.316$ for Cued recall; $g = 0.238$ for Free recall; $g = 0.336$ for a mixture of different test formats), and test format does not significantly modulate the benefits of testing.
Q4. Can knowledge tested in one format be retrieved to answer questions presented in a different format?	The classroom testing effect is significantly greater than 0 when test formats are mismatched between quizzes and exams, indicating transfer of the effect to different test formats (i.e., applying knowledge tested in one format to answer questions presented in a different format). Nevertheless, consistent test formats ($g = 0.531$) are associated with a significantly larger effect size than inconsistent formats ($g = 0.399$).
Q5. Does testing benefit untested knowledge?	Testing significantly benefits untested knowledge, although to a smaller extent ($g = 0.321$) than that for tested knowledge ($g = 0.512$).
Q6. Should corrective feedback be offered?	Offering corrective feedback following class quizzes ($g = 0.537$) significantly increases learning gains over not providing feedback ($g = 0.374$).
Q7. Does the number of test repetitions matter?	There is a positive relationship between the number of test repetitions and the classroom testing effect, indicating that the more occasions on which class content is quizzed, the larger the learning gains.
Q8. Does test-enhanced learning work at all levels of education?	Test-enhanced learning generalizes to elementary school ($g = 0.328$), middle school ($g = 0.597$), high school ($g = 0.655$), and university/college ($g = 0.486$). No firm conclusion can be made about whether it works in continuing education ($g = 0.314 [-0.081, 0.709]$, $p = .119$) because few studies ($k = 6$) have explored this.
Q9. Does testing enhance classroom learning to different extents for male and female students?	There is no reliable association between female gender ratios and the classroom testing effect, indicating that male and female students benefit from testing to a comparable extent.
Q10. Does test-enhanced learning generalize to a range of subjects?	Across 18 subject categories, testing consistently facilitates learning achievement.
Q11. Does testing benefit different levels of knowledge?	Testing is not only beneficial for learning facts ($g = 0.524$), but also promotes conceptual learning ($g = 0.644$) and facilitates knowledge application in the service of problem solving ($g = 0.453$).
Q12. Should tests be administered in or out of the classroom?	Quizzes administered in the classroom ($g = 0.514$) tend to be more beneficial than those administered out of the classroom ($g = 0.401$).
Q13. Should tests be administered pre- or postclass?	Both pre- ($g = 0.186$) and postclass ($g = 0.536$) quizzes significantly enhance learning, but postclass quizzes are more effective.
Q14. Does administration mode matter?	Administration mode (e.g., paper-and-pen, clicker response system, online) does not significantly modulate the classroom testing effect.
Q15. How does the effectiveness of test-enhanced learning vary with treatment duration?	Testing benefits systematically increase from single class treatment ($g = 0.385$), treatment lasting less than a semester ($g = 0.521$), treatment lasting a whole semester ($g = 0.547$), to that lasting longer than a semester ($g = 0.624$). There is a significantly positive relationship between the classroom testing effect and the total number of class meetings. Hence, the relationship between test-enhanced learning and treatment duration is approximately linearly increasing.
Q16. Does stake level matter?	There is no significant difference in testing benefits between high- ($g = 0.441$) and low-stake ($g = 0.477$) quizzes.
Q17. Should students take class quizzes independently or collaboratively?	Even though the results show that there is no significant difference in testing benefits between collaborative ($g = 0.653$) and independent ($g = 0.490$) quizzes, no firm conclusion can be reached at present because few studies ($k = 21$) have explored the effectiveness of collaborative testing.
Q18. What research characteristics modulate the effect size of test-enhanced learning?	Experimental design has a significant influence, with larger benefits for within-subjects designs ($g = 0.674$) than for between-subjects ones ($g = 0.415$). Instructor matching (i.e., whether the quizzed and un-quizzed students are taught by the same instructors) does not significantly modulate test-enhanced learning (Same instructor: $g = 0.459$; Different instructors: $g = 0.527$).
Q19. What are the mechanisms underlying the classroom testing effect?	The research findings support the additional exposure, transfer-appropriate processing, and motivation theories to account for the classroom testing effect, but not the retrieval effort theory (for details, see the Discussion).

meta-analysis finds that class quizzing significantly improves students' assessment performance, with a medium effect size.

Q.2 Against What Comparison Treatments Does Quizzing Enhance Learning?

The effectiveness of test-enhanced learning has been widely explored by comparing testing with a variety of other strategies, such as restudying, concept mapping, and so on. In classroom research, the most widely used strategy in the control condition is no or filler activity, wherein students either do not undertake any other activities or complete irrelevant filler tasks. Other studies have compared the effectiveness of testing with that of restudying, wherein students restudied the curriculum content or reread the textbook (or class notes). A small proportion of studies compared testing with other elaborative strategies, such as concept mapping, note-making, summarizing, and so on. Exploring whether control strategies modulate test-enhanced learning brings important theoretical implications. For instance, according to the additional exposure explanation, we expect that learning gains in the treatment (quiz) group will be larger in comparison with no or filler activity (with low reexposure in the control condition) than with restudying (with high reexposure). From a practical perspective, it is important to explore whether testing more effectively enhances classroom learning beyond other elaborative strategies (such as concept mapping, note-taking, and so on).

Q.3 Does Quiz Format Matter?

There are dozens of test formats employed in class quizzes, such as fill-in-the-blank, cued recall, free recall, short answer, matching, multiple choice, short essay, and so on. Practically, it is crucial to explore which testing formats are most efficient for enhancing classroom learning. Theoretically, exploring whether recall tests (e.g., cued recall, free recall) are more beneficial than recognition tests (e.g., multiple choice, matching) can offer a means to test the retrieval effort theory (Rowland, 2014). Recall that this account predicts larger learning gains for difficult recall tests than for easy recognition tests.

Q.4 Can Knowledge Tested in One Format Be Retrieved to Answer Questions Presented in a Different Format?

An important aspect of the transfer of the testing effect is whether knowledge tested in one format (such as short answer) can be later retrieved to answer questions presented in a different format (such as multiple choice). Exploring whether consistency between quiz and exam formats modulates the testing effect can yield important theoretical insight regarding the contributions of transfer-appropriate processing to the classroom testing effect. Recall that this account predicts greater testing benefits when quiz and exam formats are matched than mismatched, and Rowland (2014) and Adesope et al. (2017) reported inconsistent findings on this question.

Q.5 Does Testing Benefit Untested Knowledge?

The concern regarding whether test-enhanced learning transfers to untested information has received increasing attention in recent

years, with some studies documenting positive effects (e.g., Cho et al., 2017) and others finding null or even negative effects (e.g., Davis, Chan, & Wilford, 2017; Little, Storm, & Bjork, 2011). In a recent meta-analysis, Pan and Rickard (2018) found little enhancing effect on untested materials ($g = 0.16$, 95% CI $[-0.10, 0.43]$). However, Pan and Rickard only included 17 effects for untested materials, which means that the precision of their estimate is relatively low (as indicated by the wide confidence interval) and their nonsignificant enhancing effect might be a false negative resulting from low statistical power. In addition, as noted by Pan and Rickard (2018, p. 728), in some of their included studies (most of which were laboratory studies), there was no semantic relation between tested and untested information. Little et al. (2011) found that semantic coherence between tested and untested material is a key modulator of the enhancing effect on untested materials (i.e., testing promotes retention of untested/related materials but results in little enhancement for untested/unrelated materials). Hence, the lack of semantic relatedness between tested and untested materials might be another source of the nonsignificant enhancement observed in Pan and Rickard's meta-analysis, and their findings do not preclude its existence in the classroom because lecture contents usually are strongly related and textbook sections are cohesively organized.

Q.6 Should Corrective Feedback Be Offered?

As discussed above, Rowland (2014) and Adesope et al. (2017) reached inconsistent conclusions regarding whether corrective feedback adds additional value for test-enhanced learning. To resolve this puzzle, our meta-analysis includes a more comprehensive set of studies to explore the modulating role of feedback. Going beyond Rowland (2014) and Adesope et al. (2017), we directly explore this question in educational settings.

Q.7 Does the Number of Test Repetitions Matter?

The number of repeated tests on the same information (i.e., the total number of times that the same materials are repeatedly tested) has been assumed to be an important modulator of test-enhanced learning. Roediger and Karpicke (2006b), for instance, found that, on a 1-week final test, their STTT group (which studied a text once and then took three free recall tests) recalled more text materials than their SSST group (which studied the text three times and took a free recall test once). Such findings clearly indicate that a greater number of test repetitions can yield larger learning enhancement (for related findings, see Eriksson, Kalpouzos, & Nyberg, 2011; Karpicke & Roediger, 2007b, 2008; Kornell & Bjork, 2008; McDermott, 2006; Pyc & Rawson, 2009; Soderstrom, Kerr, & Bjork, 2016; Vaughn & Rawson, 2011; Wheeler & Roediger, 1992). However, strikingly, Adesope et al.'s (2017) meta-analysis found that the testing benefit was larger when there was only one ($g = 0.70$) compared with more than one ($g = 0.51$) test on the same material. Given this inconsistency, further exploration is required.

Q.8 Does Test-Enhanced Learning Work at All Levels of Education?

From a developmental perspective, it is important to explore whether test-enhanced learning generalizes to different education

levels (Roediger & Karpicke, 2006a), such as elementary school, middle school, high school, university/college, and continuing education. Even though test-enhanced learning has been soundly established for university/college students, some studies found little or even negative testing effects (i.e., testing impairs learning) for elementary students (e.g., Leahy, Hanham, & Sweller, 2015).

Q.9 Does Testing Enhance Classroom Learning to Different Extents for Male and Female Students?

Gender has a variety of influences on learning and memory (e.g., Asperholm, Högman, Rafi, & Herlitz, 2019; Herlitz, Nilsson, & Bäckman, 1997; Lewin & Herlitz, 2002). Its modulating role in the testing effect has also been repeatedly explored, but with inconsistent results. For instance, Gokcora and DePaulo (2018) observed that female students benefited more from class quizzing, whereas by contrast Nakos and Whiting (2018) observed the exact reverse pattern, with greater test-induced enhancement for male students. Given the inconsistency of the research findings, it is striking that none of the previous meta-analyses explored the modulating role of gender in the testing effect.

Q.10 Does Test-Enhanced Learning Generalize Across a Range of Subjects?

The magnitude of test-enhanced learning tends to be material-dependent (Rowland, 2014). Different types of knowledge are delivered in different subjects (e.g., Biology, History). Hence, it is critical to explore whether the classroom testing effect generalizes to different subjects. Going beyond Schwieren et al. (2017), who only examined the effects obtained in Psychology classes, the current review included effects from 31 different subjects.

Q.11 Does Testing Benefit Different Levels of Knowledge?

Retrieval practice has been met with the criticism that it is a “drill-and-kill” strategy and only enhances “inert knowledge,” which cannot be utilized to solve new problems in unfamiliar contexts (Fey, 2012). To test this claim, the current review asks whether class testing can enhance conceptual (high-level) learning and problem-solving (knowledge application to solve new problems). We coded assessment (exam) content into three categories: Fact learning, Concept learning, and Problem-solving. For low-level Fact learning, students’ main task is to remember specific facts, such as foreign words, historical events, and so on. For high-level Concept learning, students are required to comprehend information by integrating different pieces of information and make uncertain inferences that go beyond direct experience (Jacoby, Wahlheim, & Coane, 2010; Yang & Shanks, 2018). For instance, elementary students are required to learn different mathematical concepts and language students are required to master rules of syntax. For Problem-solving, students are expected to apply learnt knowledge or skills to solve new problems in new contexts.

Q.12 Should Tests Be Administered In or Out of the Classroom?

One possible reason why teachers are sometimes reluctant to administer class quizzes is that they think administering them is

time-consuming and reduces time for didactic teaching. To prevent it from occupying teaching time, some researchers have offered quizzes as homework or published the quiz questions online, with students completing them out of the classroom (e.g., Grimstad & Grabe, 2004; Marden, Ulman, Wilson, & Velan, 2013). We therefore ask whether quiz administration location (i.e., inside vs. outside the classroom) modulates test-enhanced learning? From a practical perspective, it is important to explore which type of quiz (in or out of the classroom) produces superior gains. It is surprising that, to our knowledge, this question has never been explored before.

Classroom quizzes are typically supervised by instructors and are mandatory class activities. By contrast, quizzes administered outside the classroom are usually undertaken without supervision and unproctored, and students may not fully engage in those quizzes or may even cheat in them. Hence, it is reasonable to assume that classroom quizzes may produce larger learning gains than ones administered out of the classroom. However, on the other hand, it is also possible that quizzes administered out of the classroom can produce greater benefits because they do not reduce teaching time. Overall, whether quizzes administered in or outside the classroom are more effective has not been explored before, it is difficult to make a clear prediction, and a direct investigation is needed.

Q.13 Should Tests Be Administered Pre- or Postclass?

Even though most classroom studies have administered quizzes at the end of a class or a unit wherein students were quizzed on the taught content, many other studies have explored the effectiveness of preclass quizzes, with to-be-taught content tested before formal lecturing. Preclass quizzing is expected to enhance learning by stimulating students to keep up with course readings before class and by informing them what to remember during class (Graham, 1999; Johnson & Kiviniemi, 2009; Narloch, Garbin, & Turnage, 2006). Nonetheless, there is little agreement regarding whether preclass quizzes can enhance student achievement (Gunasekera, 1997; Johnson & Kiviniemi, 2009). Hence, a meta-analysis is required to test the existence or absence of an enhancing effect of preclass quizzes. If the answer is affirmative, another important question arises: is pre- or postclass quizzing more beneficial? This question can be explored by assessing the modulating role of administration timepoint (preclass vs. postclass).

Q.14 Does Administration Mode Matter?

Although class tests are typically administered in a conventional modality (i.e., paper-and-pen), with the development of new technologies, more and more modes are available to make it easier to administer tests, such as clicker response systems, smartphones, online websites (e.g., Moodle), and personal computers. Little agreement has been reached about whether administration mode modulates the testing effect (e.g., Bojinova & Oigara, 2011; Lantz & Stawiski, 2014; Zheng & Bender, 2019).

Q.15 How Does the Effectiveness of Test-Enhanced Learning Vary With Treatment Duration?

As discussed above, the laboratory testing effect might be largely attributable to the mechanisms underlying the classic test-

ing effect on studied information; by contrast, the mechanisms proposed to account for the forward testing effect (i.e., testing enhances new learning) and those for the classic testing effect may jointly contribute to the classroom testing effect. Hence, it is reasonable to expect a larger enhancement in studies administering quizzes across multiple classes (wherein many study-test cycles are involved, with different materials studied in each cycle) than in studies administering a quiz in a single class (wherein all information is studied in a single session and there is only one study-test cycle). To evaluate this possibility, we test whether the effect size of test-enhanced learning in multiple classes is larger than that in a single class. To foreshadow, this expectation is confirmed by the current meta-analysis. Then, another important question comes to the fore: How does the effectiveness of quiz-enhanced learning vary with treatment duration?

The motivation theory yields two contrasting predictions regarding how the effectiveness of test-enhanced learning may evolve with treatment duration. The first is that the longer the treatment lasts, the more effectively quizzing will benefit student learning. Supporting evidence comes from a study by Schrank (2016). Schrank found that students' attendance in a daily exam class (which undertook a quiz each day) was maintained at a steady level across a whole semester (about 90%), whereas attendance in the control section (which did not take a daily quiz) steadily decreased from the beginning (about 90%) to the end of the semester (about 66%). This finding implies that without frequent class quizzes, student motivation and engagement are likely to decrease gradually as the course continues, but this decrease may be prevented by daily quizzes (for related findings, see Healy et al., 2017; Yang et al., 2017). Hence, it is possible that the effectiveness of test-enhanced learning systematically increases as a function of treatment duration.

The second prediction is an inverse U-shaped relation between test-enhanced learning and treatment duration. Specifically, the effect should be small for short treatment, increase from short to medium treatment, and decrease from medium to long treatment. Prior quizzes may initially motivate students to study harder and promote their attainment, but they may gradually get used to quizzing as the course continues, with the motivational consequences of quizzing decreasing across the course, leading to a smaller enhancement for long treatment. There is some support for this hypothesis (J. E. Steele, 2003). J. E. Steele taught Physiology courses across an academic year (i.e., two semesters). In the first semester, students who took class quizzes performed better on course exams than those who did not take quizzes, but test-enhanced learning was absent in the second semester (for related findings, see Golding, Wasarhaley, & Fletcher, 2012). Overall, two distinct predictions can be derived from the motivation theory about the role of treatment duration in the classroom testing effect, which are tested in the current review.

Q.16 Does Stake Level Matter?

To motivate students to learn better, some instructors increase the stake-level of class quizzes by informing students that their quiz performance will be included in their course grade (Khanna, 2015) or by offering extra awards (such as course credit) for superior quiz performance (Michaels, 2017). However, a potential side effect of increasing quiz stake is that it may induce high test

anxiety (Khanna, 2015; Tobias, 1985; Tse & Pu, 2012), which brings a variety of learning detriments (e.g., difficulty concentrating, poor test performance). Hence, it is difficult to make a clear prediction about whether increasing the stake-level of class quizzes will boost or impair learning. This question will be explored by asking whether stake level of class quizzes modulates the classroom testing effect.

Q.17 Should Students Take Class Tests Independently or Collaboratively?

Even though students typically take class tests independently, some studies have explored the effects of collaborative testing, wherein a small group of students jointly answer a set of test questions through group discussion. The effectiveness of collaborative testing is still under debate, with some studies supporting collaborative testing over independent testing (e.g., Martin, Friesen, & De Pau, 2014; Rezaei, 2015) and others showing no difference (e.g., Haberyan & Barnett, 2010; Leight, Saunders, Calkins, & Withers, 2012). The inconsistent findings will be reevaluated in the current meta-analysis.

Q.18 What Research Characteristics Modulate the Effect Size of Test-Enhanced Learning?

Different studies with different research designs and characteristics may yield different estimates of the magnitude of the classroom testing effect. Two research characteristics are evaluated in the current review. The first is whether test manipulation is implemented within-subjects (whereby some curriculum content is tested and other untested, and the testing effect is quantified by a within-subjects comparison) or between-subjects (whereby some students are tested and others untested, and the testing effect is quantified by a between-subjects comparison). The second is whether the courses in the treatment and control conditions are delivered by the same instructor(s).

Q.19 What Are the Mechanisms Underlying the Classroom Testing Effect?

As discussed above, the laboratory and classroom testing effects may be driven by different mechanisms, and the classroom testing effect's underlying mechanisms have scarcely been explored. Hence, an important goal of the current meta-analysis is to investigate its cognitive and motivational underpinnings.

Method

Literature Search

To obtain a comprehensive set of eligible studies, we conducted a systematic search using the following search term: [testing effect* OR test-enhanced learning OR test* OR retrieval practice* OR quiz* OR exam* OR assessment*] AND [classroom* OR class* OR lecture* OR course*]. The search was performed in the following electronic databases: Web of Science, PubMed, PsycARTICLES, PsycINFO, Google Scholar, and the ProQuest Dissertation & Theses Global Database. In a pilot search before initiating this project, Google Scholar returned over 421,000 articles. Be-

cause Google Scholar only allowed access to the first 1,000 results, we preplanned to only screen this number of articles. In addition, we used different combinations of the above search terms to reconduct the study search several times to ensure comprehensiveness.

We took further steps to identify eligible studies. The reference lists of the seven meta-analytic reviews cited previously and their Google Scholar citations were screened. The 1,965 Google Scholar citations of Roediger and Karpicke (2006a) were also manually checked. We emailed 134 corresponding authors, whose studies were identified as eligible in the above search and which were published after (and including) the year 2000, to request any unpublished studies meeting the inclusion criteria, and they were also encouraged to forward our e-mail request to any other researchers who might have relevant data.

Inclusion and Exclusion Criteria

1. Only classroom studies were included, for which at least the initial learning phase (e.g., lecturing) occurred in the classroom. In other words, curriculum content had to be delivered in the classroom. In some studies, quizzes were administered outside the classroom (e.g., online) with the aim of freeing up teaching time (e.g., Grimstad & Grabe, 2004). Such studies were included to explore the moderating effect of quiz administration location, as discussed above. Laboratory and online education (distance education) studies were not included.
2. Only studies which compared testing/quizzing with no quizzing (or quizzing with fewer test questions) were included. Some studies explored the test frequency effect (e.g., Kling, McCorkle, Miller, & Reardon, 2005; Murphy & Stanga, 1994), with some students taking short quizzes frequently (e.g., each day) and others taking long quizzes less frequently (e.g., each week or month). In these studies, the frequently and less frequently quizzed students were tested on the same number of questions and the same materials, except that the test frequency was different. Such studies were excluded because all students were tested on the same questions and the same materials. Readers interested in the test frequency effect can consult Bangert-Drowns et al. (1991) for a comprehensive review.
3. Duplicates were excluded. In addition, if the same results were reported in both a thesis and a journal article published by the same authors (e.g., Carpenter, Rahman, & Perkins, 2018; Rahman, 2017), the thesis was excluded.
4. Empirical studies were included. Qualitative interviews, questionnaire survey studies, studies only involving subjective measures (e.g., *How well do you think you have mastered the knowledge in this chapter?*) without objective measures (e.g., recall tests) of learning outcomes, and review articles were not included.
5. Only studies reporting sufficient information for effect size calculation were included.

6. Only articles written in English were considered.

The screening procedure and results are reported in a flowchart (see Figure 3).

Data Extraction and Analysis

The first author (CY) and a research assistant independently performed data extraction and moderator coding. The research assistant was trained on how to perform the data extraction and moderator coding before the project commenced. All divergences were settled through discussion.

If Cohen's d s were reported in the original reports, we directly extracted the reported values. Otherwise, Cohen's d s were calculated using the formulae provided by Borenstein, Hedges, Higgins, and Rothstein (2009). For within-subjects effects, correlations between exam performance in the treatment and control conditions are required to adjust within-group standard deviations (SD s) and Hedge's g s. Unfortunately, few studies reported these values. However, 59 within-subjects effects, identified in the current meta-analysis, simultaneously reported means, SD s (or SE s), and paired-samples t values (or Cohen's d s). These data enabled us to calculate the correlations (r s) between their dependent measures (Morris & DeShon, 2002). The 59 r coefficients were transformed into Fisher's Z scores (Silver & Dunlap, 1987) and then submitted to a multilevel random-effects meta-analysis, which found a significantly positive correlation between dependent measures, $Z = 0.494$ [0.370, 0.618], $p < .001$. This Z score was then transformed back to $r = .457$ [0.354, 0.550]. Accordingly, the current meta-analysis imputed a correlation of 0.457 for all within-subjects design effects.²

To mitigate potential bias in effects with small sample sizes, all Cohen's d s were transformed into Hedge's g s using the bias correction function provided by Hedges (1982). Given that some effects were extracted from a single study (and the same sample), which might violate the assumption of independence, all meta-analyses were performed using multilevel random-effects models (except where noted otherwise), adding a random intercept at the study level (Pastor & Lazowski, 2018; Van Den Noortgate & Onghena, 2003). All analyses were conducted via the R *metafor* package (Viechtbauer, 2010) unless noted otherwise.

Results

The database search procedure identified 220 studies as eligible, while correspondence with researchers elicited two more research projects.³ From those 222 projects (marked by * in the reference list), 573 effects and 48,478 students' data were extracted.

² Another approach to address the correlation issue is to follow the recommendation from Cumming (2012) by imputing a correlation of 0.500 for all within-subjects design effects, as was done in previous meta-analyses (e.g., Chan, Meissner, & Davis, 2018; Pan & Rickard, 2018; Rowland, 2014). We note that, regardless of setting r to 0.457 or 0.500, all results showed the same patterns.

³ We thank Antônio Jaeger and Autumn B. Hostetter for sharing their unpublished data with us.

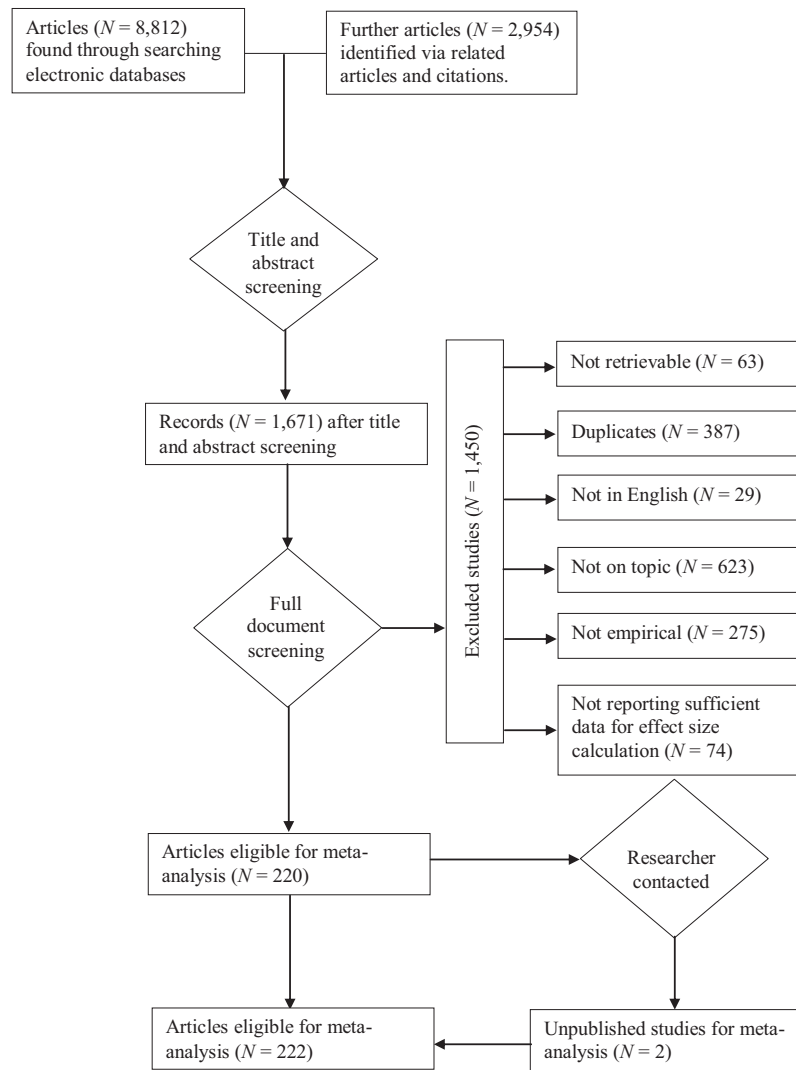


Figure 3. Flowchart depicting the article screening results.

Test-Enhanced Learning in the Classroom

The weighted mean effect size, generated by a multilevel random-effects model, was $g = 0.499$ [0.442, 0.557], $p < .001$, indicating that quizzing raises academic achievement scores overall by approximately one half standard deviation. Heterogeneity among the effects was substantial, $Q(572) = 4,816$, $p < .001$. The median effect size was $g = 0.446$, slightly smaller than the weighted mean. As shown in Figure 4, the overall distribution of effects was right skewed (skewness value = 1.096). The majority (82.9%) of effects were positive, with a minority (15.5%) negative and only 1.6% showing a null effect ($g = 0$).

Outlier and Influential Case Diagnostics

Sensitivity analyses (such as leave-one-out, Cook's distances) were conducted to detect outliers and influential cases. Because *metafor* does not apply these analyses to multilevel meta-analysis, we ran them using a conventional random-effects model. The

results identified seven outliers (with g s ranging from 2.234 to 3.269). After removing these seven effects, we reran the multilevel random-effects meta-analysis, which found that the weighted mean fell slightly to $g = 0.476$ [0.425, 0.527], $p < .001$, $Q(565) = 4,066$, $p < .001$. Viechtbauer (2010) noted that the detection of an outlying/influential effect does not automatically merit its deletion. In addition, when we rechecked these effects, we did not find any miscoding. Hence, we decided to retain them in the following analyses, and we note that including or excluding these seven effects does not significantly affect the overall result patterns.

Publication Bias: p -Curve

p -curve analysis is a recently developed measurement tool to detect the existence or absence of an effect by evaluating the distribution of significant p values (Simonsohn, Nelson, & Simmons, 2014a). For a true effect, the distribution of significant p values (i.e., $p < .05$) should be significantly right-skewed, with $p < .025$ much more prevalent than $.025 < p < .05$. For a spurious

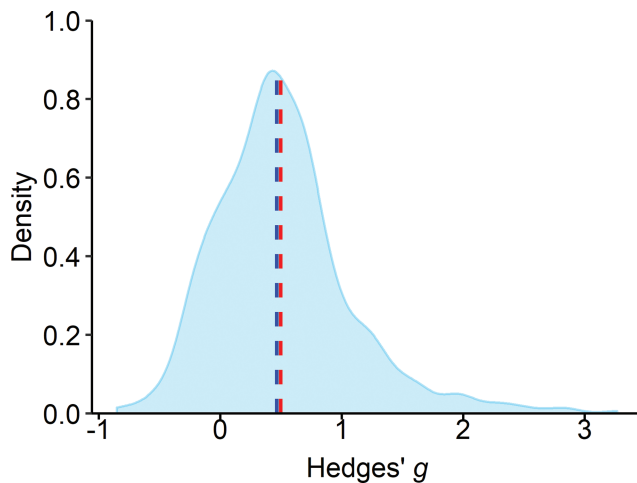


Figure 4. Kernel density distribution plot depicting the density distribution of the included effects. The dotted red and blue lines represent the weighted mean and median, respectively. See the online article for the color version of this figure.

effect without publication bias and/or *p*-hacking, all significant *p* values should appear with equal probability and the distribution of significant *p* values should be flat. For a spurious effect with publication bias/*p*-hacking, the distribution of significant *p* values would be left-skewed.

Following Simonsohn et al.'s guidelines (available at <http://www.p-curve.com/guide.pdf>; also see Simonsohn, Nelson, & Simmons, 2014b), the corresponding *p* values were selected from the key contrasts. *p*-curve analysis was conducted via the online application (Version 4.06) developed by Simonsohn and colleagues (available at <http://www.p-curve.com/app4/>), and the corresponding distribution is depicted in Figure 5. The right skewness of the *p*-curve was significant, $z = -36.85$, $p < .001$ (full *p*-curve), and $z = -34.80$, $p < .001$ (half *p*-curve), confirming that the included studies contain evidential value for the existence of the classroom testing effect. In addition, the *p*-curve did not indicate evidential inadequacy (i.e., flatter than 33% power), $z = 25.64$, $p > .999$ (full *p*-curve), and $z = 32.19$, $p > .999$ (half *p*-curve). The estimated power of tests included in the *p*-curve was 99%. Overall, the *p*-curve analysis finds strong evidence for the existence of test-enhanced learning in the classroom.

Publication Bias: Other Methods

Five quantitative methods were employed to detect whether the included effects were biased. Before reporting the results, we highlight that there is currently no perfect technique for assessing and correcting publication bias (Pham, Platt, McAuley, Klassen, & Moher, 2001; Stanley, 2017), and existing methods suffer from various limitations and weaknesses (Carter, Schönbrodt, Gervais, & Hilgard, 2019). For instance, PET-PEESE (see below) often overadjusts the effects leading to underestimation of effect size (Gervais, 2015); Duval and Tweedie's *Trim-and-Fill* method is known to undercorrect for publication bias (Hilgard, 2017); although publication year and status are possible indices of publication bias, little can be done to correct them; and the model

selection approach may suffer from convergence problems, especially when the number of included effects is small (Terrin, Schmid, Lau, & Olkin, 2003). We hence applied a variety of widely used methods to provide a comprehensive assessment of potential publication bias in the included effects.

The first is to explore the relationship between the magnitude of the effect and publication year. The logic behind this analysis is that, if initial results are biased by selective publication of significant findings, later studies addressing the same finding are unlikely to obtain equally large effects. Therefore, for a spurious effect, reported effect sizes ought to gradually decrease across years (Borenstein & Cooper, 2009). A multilevel random-effects metaregression, regressing *g*s onto publication year (ranging from 1929 to 2019), showed no relationship between these two variables, $\beta = 0.0006$, $Q(1) = 0.091$, $p = .763$ (see Figure 6), indicating little risk of time-based publication bias.

The second method is to explore whether publication status modulates the effects. The logic of this method is that significant results (with larger effect sizes) are more likely to be published than nonsignificant ones (with smaller effect sizes). We coded the 573 effects into two categories based on their publication status: Published ($k = 487$, including 485 from journal articles and 2 from book chapters) and Unpublished ($k = 86$, including 61 from dissertations, 21 from conference reports, and 4 from unpublished projects). A multilevel metaregression analysis showed that publication status did not significantly modulate the included effects, $Q(1) = 0.607$, $p = .436$, with $g = 0.510$ [0.447, 0.573], $p < .001$, for published effects and $g = 0.449$ [0.311, 0.588], $p < .001$, for unpublished effects. Again, these results reveal minimal evidence of publication bias.

The third method is Egger's regression, which assesses the relationship between *g*s and their corresponding *SE*s (Egger, Smith, Schneider, & Minder, 1997), shown in the funnel plot in Figure 7. Considering the limitations of Egger's regression, PET-PEESE was employed (Stanley & Doucouliagos, 2014). To our

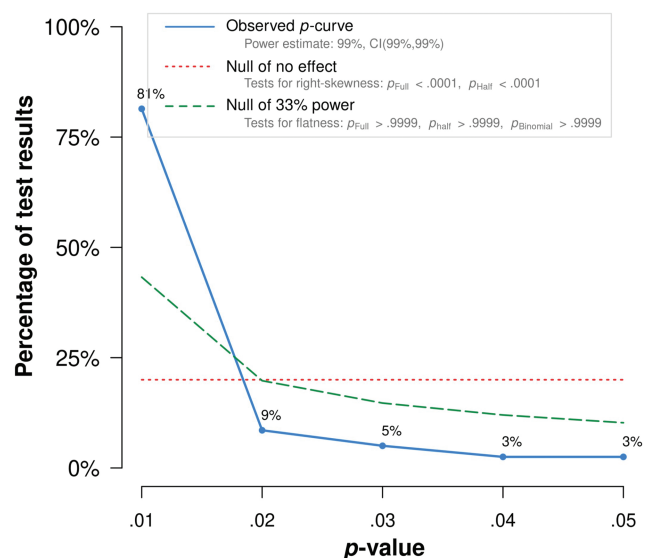


Figure 5. *p*-curve plot depicting the distribution of significant *p* values. See the online article for the color version of this figure.

knowledge, PET-PEESE is not applicable to multilevel meta-analysis in *metafor* and we therefore performed PET and PESEE adjustments with robust variance estimation, a method for estimating the dependence of effects within studies (Hedges, Tipton, & Johnson, 2010). The robust variance estimation analyses were performed via the *R robumeta* package (Fisher & Tipton, 2015). Because both PET and PESEE analyses showed that the regression intercept was significantly greater than 0, the PESEE test was taken to assess bias and to estimate the corrected intercept, $g = 0.475$ [0.329, 0.620], $p < .001$, and the funnel plot was not significantly asymmetric, $t(5.17) = 0.354$, $p = .737$, indicating little risk of publication bias.

The fourth method is Duval and Tweedie's *Trim-and-Fill* method (Duval & Tweedie, 2000). This method gradually removes or "trims" the effects with large *SEs* until the funnel plot is symmetric, after which the removed studies and their missing counterparts around the center are "filled" back into the funnel plot to maintain its symmetry. Because the Trim-and-Fill analysis is not implemented for multilevel meta-analysis in *metafor* and is incompatible with robust variance estimation, we employed a conventional random-effects model. The result showed that the estimated number of missing effects detected by this method was 0, again implying little need to worry about publication bias.

The fifth method is to apply a three-parameter selection model (3PSM; McShane, Böckenholt, & Hansen, 2016; Vevea & Hedges, 1995; Vevea & Woods, 2005), which considers three parameters to assess and correct publication bias: an effect size parameter (i.e., the weighted average of effects), a heterogeneity parameter (i.e., the degree of heterogeneity among effects),

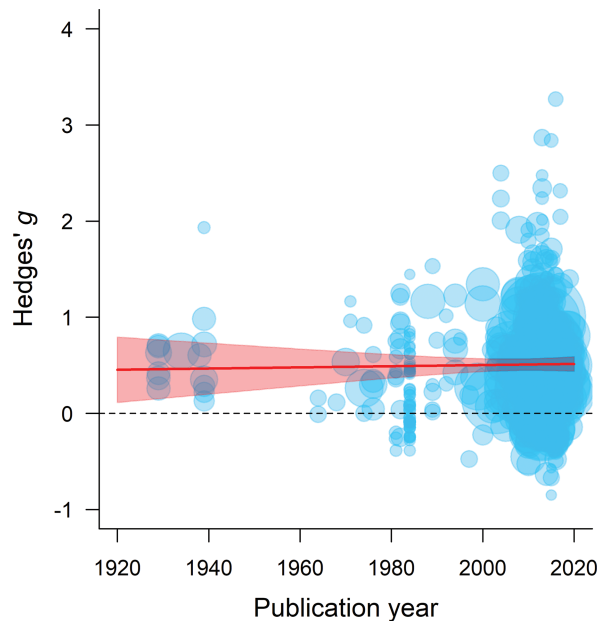


Figure 6. Bubble plot depicting the relationship between publication year and the reported effect sizes of the classroom testing effect. Bubble sizes represent the relative weights of the effects, and error bars represent the 95% CI of the regression trend. See the online article for the color version of this figure.

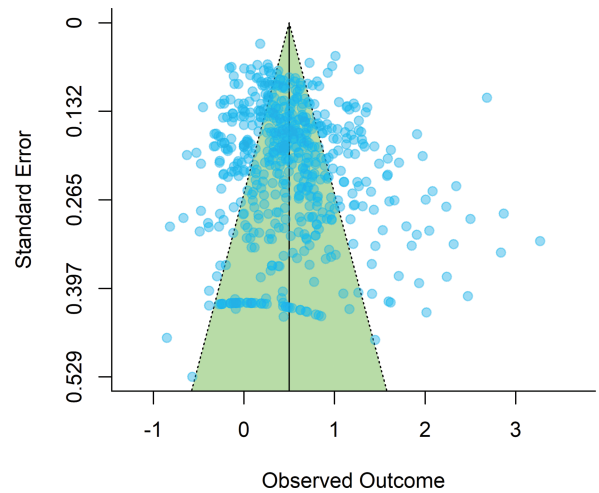


Figure 7. Funnel plot depicting the relationship between *gs* and standard errors. The vertical line on which the funnel is centered represents the weighted mean effect size, and the light-green zone represents the 95% CI of the weighted mean. See the online article for the color version of this figure.

and a selection parameter (i.e., the relative ratio of the likelihood that insignificant results are published over the likelihood that significant results are published). 3PSM has been shown to be more reliable than other conventional bias assessment methods (e.g., McShane et al., 2016; Pustejovsky & Rodgers, 2019). We had to apply 3PSM to a conventional random-effects meta-analysis because it is again not compatible with multilevel meta-analysis or robust variance estimation. This analysis was conducted via the *R weightr* package (Coburn & Vevea, 2019) and yielded no statistically significant evidence of publication bias, $\chi^2(1) = 2.999$, $p = .083$, although hinting at bias slightly more strongly than the other methods. The adjusted effect was $g = 0.461$ [0.390, 0.532], $p < .001$.⁴

In summary, five quantitative methods were employed to detect potential publication bias in the included effects, and the results consistently point to little risk of publication bias in the current review.

Moderator Analyses

To unravel the moderators and boundary conditions of the classroom testing effect, metaregression analyses were conducted. Categorical and continuous moderator analysis results are summarized in Tables 2 and 3, respectively.

Control strategy. Study strategies employed in the control condition were coded into four categories: No/Filler activity (students either completed no further activities or irrelevant filler tasks

⁴ Another approach to implement 3PSM is to use bootstrapping (Hilgard, Sala, Boot, & Simons, 2019). Specifically, we randomly selected an effect from each study and applied 3PSM to obtain an estimated *g*. This process was repeated 1,000 times to obtain 1,000 *gs*. Finally, the median of the bootstrapped *gs* was treated as the corrected *g*, and its 95% CI were defined as the 2.5th and 97.5th percentile of the bootstrapped estimates. Overall, the adjusted effect was $g = 0.427$ [0.395, 0.458].

Table 2
Categorical Moderator Analysis Results

Categorical modulators	<i>k</i>	<i>g</i>	95% CI	<i>Q_B</i>	<i>p</i>
Publication status				0.607	.436
Published	487	0.510	[0.447, 0.573]		<.001
Unpublished	86	0.449	[0.311, 0.588]		<.001
Control strategy				153	<.001
No/filler activity	345	0.610	[0.547, 0.673]		<.001
Fewer questions	25	0.465	[0.218, 0.711]		<.001
Restudying	162	0.330	[0.256, 0.404]		<.001
Elaborative strategies	41	0.095	[−0.005, 0.194]		.062
Quiz format				61.333	<.001
Matching	6	0.913	[0.691, 1.135]		<.001
Fill-in-the-blank	22	0.773	[0.578, 0.969]		<.001
Short answer	93	0.638	[0.539, 0.737]		<.001
Multiple choice	270	0.567	[0.496, 0.638]		<.001
Cued recall	20	0.316	[0.124, 0.509]		.001
Free recall	22	0.238	[0.082, 0.394]		.003
Mixed	80	0.336	[0.236, 0.435]		<.001
Others ^a	4	0.252	[−0.197, 0.702]		.271
Unknown	56	0.447	[0.326, 0.568]		<.001
Recognition versus recall				0.004	.952
Recall	157	0.520	[0.431, 0.608]		<.001
Recognition	278	0.518	[0.438, 0.597]		<.001
Recognition versus recall (with feedback)				1.583	.208
Recall	97	0.541	[0.422, 0.660]		<.001
Recognition	165	0.607	[0.504, 0.710]		<.001
Test format matching				23.378	<.001
Yes	328	0.531	[0.464, 0.599]		<.001
No	151	0.399	[0.325, 0.473]		<.001
Unknown	94	0.532	[0.426, 0.638]		<.001
Test material matching				22.367	<.001
Identical	226	0.512	[0.432, 0.591]		<.001
Rephrasing	100	0.558	[0.464, 0.651]		<.001
Partial	41	0.496	[0.363, 0.628]		<.001
Untested	43	0.321	[0.210, 0.432]		<.001
Unknown	163	0.491	[0.402, 0.579]		<.001
Corrective feedback				11.164	.004
Yes	335	0.537	[0.469, 0.604]		<.001
No	142	0.374	[0.278, 0.471]		<.001
Unknown	96	0.500	[0.382, 0.618]		<.001
Test repetition				54.995	<.001
1	391	0.444	[0.383, 0.506]		<.001
2	76	0.601	[0.502, 0.700]		<.001
≥ 3	78	0.642	[0.542, 0.742]		<.001
Unlimited	28	0.762	[0.652, 0.873]		<.001
School level				6.848	.144
Elementary school	43	0.328	[0.085, 0.571]		.008
Middle school	140	0.597	[0.428, 0.765]		<.001
High school	49	0.655	[0.496, 0.815]		<.001
University/college	335	0.486	[0.420, 0.552]		<.001
Continuing education	6	0.314	[−0.081, 0.709]		.119
Subject				13.642	.692
Accounting/business/economics/finance	16	0.313	[0.004, 0.622]		.047
Biology	20	0.409	[0.135, 0.683]		.003
Chemistry	10	0.340	[0.027, 0.654]		.034
Ecology	6	0.660	[0.276, 1.045]		<.001
Education	8	0.745	[0.337, 1.154]		<.001
Engineering	11	0.584	[0.240, 0.928]		<.001
General science ^b	125	0.531	[0.337, 0.725]		<.001
Geography	9	0.690	[0.409, 0.971]		<.001
History	24	0.625	[0.374, 0.876]		<.001
Language/reading/vocabulary	55	0.644	[0.447, 0.841]		<.001
Material science	8	0.672	[0.083, 1.261]		.025
Mathematics/statistics	27	0.433	[0.193, 0.674]		<.001
Medical	39	0.481	[0.309, 0.653]		<.001
Nursing	7	0.693	[0.290, 1.096]		<.001

(table continues)

Table 2 (continued)

Categorical moderators	<i>k</i>	<i>g</i>	95% CI	<i>Q_B</i>	<i>p</i>
Pharmacy	6	0.338	[−0.053, 0.728]	8.125	.090
Physiology	41	0.396	[0.202, 0.590]		<.001
Psychology	144	0.483	[0.375, 0.591]		<.001
Others ^c	17	0.582	[0.311, 0.853]		<.001
Exam content type				10.103	.087
Fact	337	0.524	[0.453, 0.595]		<.001
Concept	91	0.644	[0.507, 0.781]		<.001
Problem-solving	59	0.453	[0.309, 0.596]		<.001
Mixed	62	0.424	[0.285, 0.563]		<.001
Unknown	24	0.350	[0.142, 0.558]		.001
Administration location				19.554	.006
In the classroom	486	0.514	[0.452, 0.576]		<.001
Outside the classroom	78	0.401	[0.290, 0.512]		<.001
Unknown	9	0.777	[0.568, 0.986]	4.297	<.001
Administration timepoint					<.001
Preclass	34	0.186	[0.036, 0.337]		.015
Postclass	487	0.536	[0.476, 0.597]		<.001
Combination	52	0.453	[0.319, 0.588]	6.728	<.001
Administration mode					.637
Simulation	6	0.641	[0.259, 1.023]		.001
Computer	25	0.590	[0.286, 0.893]		<.001
Clicker response system	85	0.549	[0.438, 0.659]		<.001
Paper-and-pen	350	0.494	[0.426, 0.562]		<.001
Oral	13	0.452	[0.276, 0.628]		<.001
Web-based	75	0.447	[0.335, 0.559]		<.001
Unknown	19	0.552	[0.391, 0.712]	9.174	<.001
Single versus multiple classes					.010
Single class	195	0.389	[0.287, 0.490]		<.001
Multiple classes	378	0.541	[0.476, 0.606]	3.233	<.001
Treatment duration					.027
Single class	195	0.385	[0.283, 0.488]		<.001
< Semester	128	0.521	[0.412, 0.631]		<.001
= Semester	234	0.547	[0.465, 0.629]		<.001
> Semester	16	0.624	[0.500, 0.748]	3.068	<.001
Stake					.199
High	160	0.441	[0.350, 0.531]		<.001
Low	135	0.477	[0.380, 0.574]		<.001
Unknown	278	0.548	[0.469, 0.626]	36.037	<.001
Collaboration					.080
Independent	552	0.490	[0.432, 0.548]		<.001
Collaborative	21	0.653	[0.472, 0.835]	9.813	<.001
Experimental design					<.001
Within-subjects	227	0.674	[0.592, 0.757]		<.001
Between-subjects	346	0.415	[0.348, 0.481]	.007	<.001
Instructor matching					.007
Same	469	0.459	[0.396, 0.523]		<.001
Different	29	0.527	[0.308, 0.747]		<.001
Unknown	75	0.670	[0.547, 0.793]		<.001

Note. Q_B represents heterogeneity for between-levels moderator tests.

^a Others includes Old/new recognition ($k = 1$), True/False judgment ($k = 1$), and Short essay ($k = 2$). ^b In some studies (e.g., McDaniel, Agarwal, Huelser, McDermott, & Roediger, 2011), elementary, middle, and high school students studied general science topics (such as water science, sun, food science, the greenhouse effect, etc.). Such courses were classified into the General science category. ^c Others include eight subjects: Aerospace ($k = 1$), Computer Science ($k = 2$), Communication ($k = 3$), Genetics ($k = 1$), Law ($k = 3$), Physics ($k = 4$), Politics ($k = 1$), and Research Methods ($k = 2$).

after class), Restudying (they restudied lecture content or reread textbooks, class notes, or lecture summaries), Fewer questions (students in the control condition were tested on fewer questions compared with those in the treatment condition), and Elaborative strategies (other study strategies).⁵ Control strategy significantly modulated the effect, $Q(3) = 153$, $p < .001$, with the largest enhancement against No/Filler activity ($g = 0.610$ [0.547, 0.673], $p < .001$) and smallest against Elaborative strategies ($g = 0.095$ [−0.005, 0.194], $p = .062$). Further analyses revealed that quiz-zing produced larger learning gains when compared with No/Filler

activity than when compared with Restudying ($g = 0.330$ [0.256, 0.404], $p < .001$), $Q(1) = 101$, $p < .001$, which is consistent with the additional exposure explanation. However, it is worth highlighting that additional exposure cannot fully explain the class-

⁵ Elaborative strategies included many varieties, such as concept mapping, note-taking, summarizing, and so on. These strategies had few effect sizes and hence were combined into a single category. We term this category Elaborative strategies because they generally involve more elaborative processing than passive restudying.

Table 3
Continuous Moderator Analysis Results

Continuous metaregression models	β	95% CI	Q_B	p
Linear model of publication year			0.091	.763
Intercept	-0.661	[-8.228, 6.907]		.864
Publication year	0.0006	[-0.0032, 0.0043]		.763
Linear model of test repetition			17.310	<.001
Intercept	0.390	[0.310, 0.469]		<.001
Test repetition	0.083	[0.044, 0.122]		<.001
Linear model of female gender ratio			0.321	.571
Intercept	0.533	[0.267, 0.798]		<.001
Female gender ratio	-0.115	[-0.515, 0.284]		.571
Linear model of NCM			5.899	.015
Intercept	0.417	[0.316, 0.518]		<.001
NCM	0.008	[0.002, 0.014]		.015
Quadratic model of NCM			6.860	.032
Intercept	0.446	[0.324, 0.567]		<.001
NCM	-0.002	[-0.023, 0.019]		.880
NCM ²	0.0003	[-0.0003, 0.0008]		.343

Note. Q_B represents heterogeneity for between-levels moderator tests. NCM = number of class meetings.

room testing effect because class quizzing more effectively facilitated learning than restudying.

Test format in quizzes. The included effects came from different test formats in quizzes which were classified into nine categories (see Table 2). The modulating effect of test format was significant, $Q(8) = 61.333, p < .001$, with Matching producing the largest benefit ($g = 0.913 [0.691, 1.135], p < .001$) and Free recall generating the smallest ($g = 0.238 [0.082, 0.394], p = .003$). Importantly, Multiple-choice tests produced significant enhancement ($g = 0.567 [0.496, 0.638], p < .001$), mitigating the concern about their effectiveness (Fazio, Agarwal, Marsh, & Roediger, 2010; Roediger & Marsh, 2005).

Further analyses were conducted to test the main prediction of the retrieval effort theory: difficult recall tests should yield larger enhancement than easy recognition ones. One hundred thirty-eight effects were excluded from the following analyses because they either involved a mixture of different formats ($k = 80$), did not report quiz format ($k = 56$), or the reported formats were short essay ($k = 2$). The remaining 435 effects were grouped into two categories: (a) Recall ($k = 157$, including 93 for short answer, 22 for free recall, 22 for fill-in-the-blank, and 20 for cued recall) and (b) Recognition ($k = 278$, including 270 for multiple choice, six for matching, one for true/false judgment, and one for old/new recognition). Inconsistent with the retrieval effort theory, Recognition tests ($g = 0.518 [0.438, 0.597], p < .001$) produced equivalent learning gain as Recall tests ($g = 0.520 [0.431, 0.608], p < .001$), $Q(1) = 0.004, p = .952$. Although this result does not support the retrieval effort theory, it should be interpreted with caution because, as noted by Rowland (2014), feedback might be important to observe larger benefits for recall tests (Kang et al., 2007).

Recall tests are generally more difficult and associated with greater recall failures than recognition tests. Hence, if corrective feedback is not provided, recognition tests in the acquisition phase may yield superior quiz performance and greater reexposure to successfully identified items than recall tests. Greater reexposure induced by recognition tests (based on the additional exposure theory) and more effective memory consolidation induced by

recall tests (based on the retrieval effort theory) may cancel out, leading to an overall null difference between recognition and recall tests.

Following Rowland (2014), the above analyses were reperformed, excluding 173 effects from studies that did not provide feedback or did not report whether feedback was provided. In total, 262 effects provided feedback following quizzing and were eligible for the following analysis. Offering feedback matched reexposure between Recall and Recognition tests, and hence provided an opportunity to directly test the retrieval effort theory. The results again found no significant difference between Recognition ($g = 0.607 [0.504, 0.710], p < .001$) and Recall tests ($g = 0.541 [0.422, 0.660], p < .001$), $Q(1) = 1.583, p = .208$, providing little support for the retrieval effort theory. Overall, the above results are inconsistent with the retrieval effort theory's explanation of the classroom testing effect. We consider these results further in the Discussion.

Test format matching. The transfer-appropriate processing theory predicts a smaller benefit for mismatched than for matched test formats. To test this hypothesis, effects were coded into three categories based on test format consistency between quizzes and exams: Yes (matched), No (mismatched), and Unknown (quiz or exam formats were not explicitly reported), and a multilevel random-effects metaregression analysis was conducted. The result revealed a significant modulating effect of format matching, $Q(2) = 23.378, p < .001$. A further analysis, in which the Unknown effects were excluded, found that matched format ($g = 0.531 [0.464, 0.599], p < .001$) was associated with larger learning gains than mismatched format ($g = 0.399 [0.325, 0.473], p < .001$), $Q(1) = 19.633, p < .001$. This outcome, which is consistent with what Adesope et al. (2017) observed but inconsistent with the findings of Rowland (2014), supports the transfer-appropriate processing theory's explanation of the classroom testing effect.

Test material matching. Effects were coded into five categories according to whether the materials in the initial test (quiz) and final assessment (exam) were identical: Identical (the same materials), Rephrasing (the same materials were tested in the quizzes and exams, while test questions were rephrased in the

exam), Partial (both quizzed and nonquizzed materials were included in the exam, but the reports did not provide sufficient information to calculate separate effect sizes for quizzed and nonquizzed materials, respectively), Untested (materials were not tested in the quizzes but tested in the exams), and Unknown (not reported). Material matching significantly modulated the testing effect, $Q(4) = 22.367, p < .001$. The significant effect on rephrased questions ($g = 0.558 [0.464, 0.651], p < .001$) demonstrates its transferability to question rephrasing. Importantly, the enhancing effect on untested materials ($g = 0.321 [0.210, 0.432], p < .001$) confirms its transfer to untested material: class quizzing not only benefits retention of tested but also retention of untested materials.

Corrective feedback. According to whether corrective feedback was offered or not, the effects were assigned into three categories: Yes (feedback provided), No (feedback not provided), and Unknown (not reported). Recall that Rowland (2014) found that corrective feedback doubles the benefits of testing but Adesope et al. (2017) observed that corrective feedback does not add any additional value. A multilevel random-effects meta-regression analysis found that corrective feedback significantly modulated the classroom testing effect, $Q(2) = 11.164, p = .004$. Consistent with Rowland (2014), offering corrective feedback following quizzes ($g = 0.537 [0.469, 0.604], p < .001$) boosted the enhancement compared with not providing feedback ($g = 0.374 [0.278, 0.471], p < .001$), $Q(1) = 10.849, p = .001$, justifying the provision of corrective feedback following class quizzes. This finding is also consistent with the additional exposure theory: Corrective feedback induces greater reexposure and larger learning gains.

Test repetition. The numbers of test repetitions (i.e., how many times the same information was repeatedly tested in the quizzes) were assigned into four categories: 1, 2, ≥ 3 , and Unlimited.⁶ The testing effect differed significantly across these four categories, $Q(3) = 54.995, p < .001$. Importantly, learning gains increased across items quizzed once ($g = 0.444 [0.383, 0.506], p < .001$), twice ($g = 0.601 [0.502, 0.700], p < .001$), and three times or more ($g = 0.642 [0.542, 0.742], p < .001$). Quizzes that permitted unlimited attempts also yielded significant enhancement, $g = 0.762 [0.652, 0.873], p < .001$.

To further explore the relationship between test repetition and the testing effect, a continuous multilevel meta-regression analysis was conducted, in which g s were regressed onto the number of test repetitions (ranging from 1–6). The Unlimited effects ($k = 28$) were excluded from this analysis. The detailed results are reported in Table 3 and visually depicted in Figure 8. There was a positive relationship between these two variables, $\beta = 0.083 [0.044, 0.122], Q(1) = 17.310, p < .001$, indicating that every additional test increased g by 0.083.

Overall, these results confirm that the more occasions on which class content is quizzed, the larger the learning gains. Although this finding is inconsistent with what Adesope et al. (2017) found, it is in line with what many empirical studies have documented (e.g., Roediger & Karpicke, 2006b).

School level. School level was coded into five categories: Elementary school, Middle school, High school, University/College, and Continuing education.⁷ The magnitude of the testing effect did not vary significantly across school levels, $Q(4) = 6.848, p = .144$. Testing produced significant enhancement in

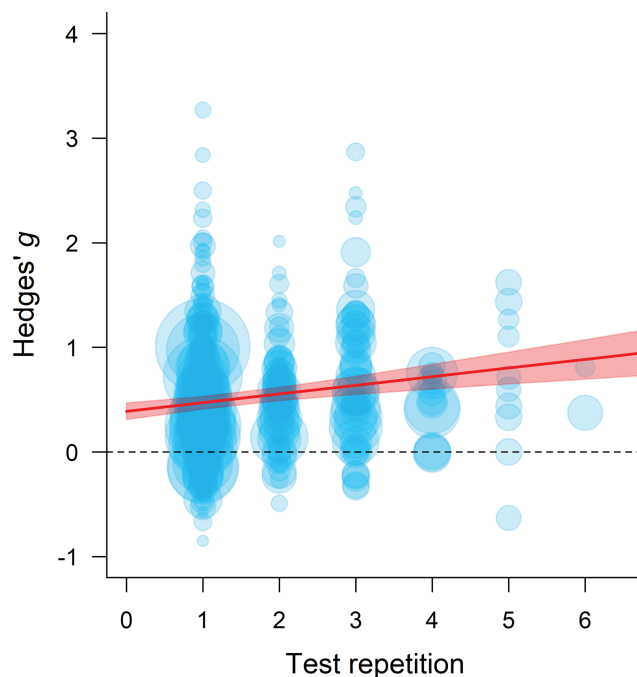


Figure 8. Bubble plot depicting the relationship between test repetition and the testing effect. Bubble sizes represent the relative weights of the effects and error bars represent the 95% CI of the regression trend. See the online article for the color version of this figure.

elementary school ($g = 0.328 [0.085, 0.571], p = .008$), middle school ($g = 0.597 [0.428, 0.765], p < .001$), high school ($g = 0.655 [0.496, 0.815], p < .001$), and university/college ($g = 0.486 [0.420, 0.552], p < .001$). Testing also numerically (although nonsignificantly) enhanced learning outcomes in continuing education ($g = 0.314 [-0.081, 0.709], p = .119$). We further assess the nonsignificant enhancement for continuing education in the Discussion.

Gender. To explore the modulating role of gender, a continuous multilevel random-effects meta-regression analysis was conducted, in which g s were regressed onto the reported female ratio (i.e., the proportion of female students contributing to each effect, ranging from 0 to 1). In total, 211 effects were included in this analysis and the other 362 were excluded for not reporting gender information. The results revealed no significant relationship between these two variables, $\beta = -0.115 [-0.515, 0.284], Q(1) = 0.321, p = 0.571$ (see Figure 9), indicating little role of gender in the classroom testing effect. Stated differently, male and female students benefited from testing to a comparable extent.

Subject. The classroom testing effect has been explored in 31 different academic subjects. Given that some subjects had a small number (fewer than five) of effects, which might result in low statistical power to produce reliable results, we combined several

⁶ In some studies (e.g., Trumbo, Leiting, McDaniel, & Hodge, 2016), quiz questions were made available to students online and they were allowed unlimited attempts to answer the test questions.

⁷ Several studies explored the testing effect on professionals attending academic workshops for continuing education (e.g., McConnell, Hou, Panju, Panju, & Azzam, 2018).

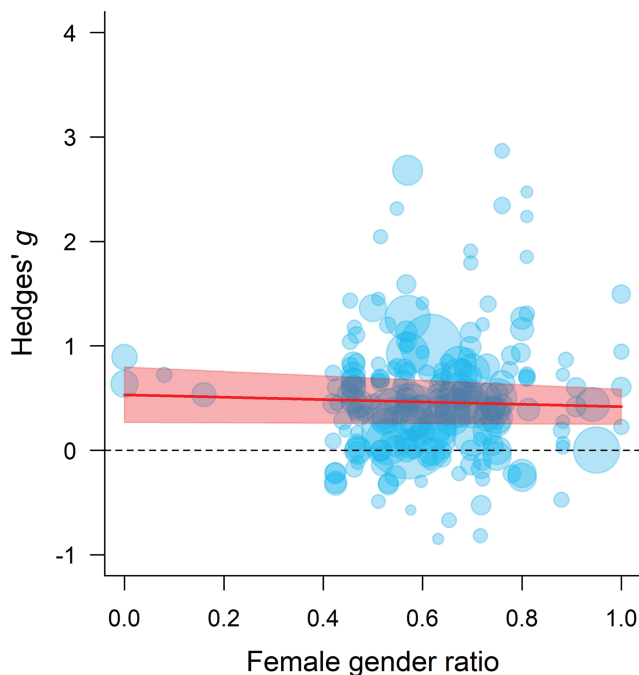


Figure 9. Bubble plot depicting the relationship between female gender ratio and the testing effect. Bubble sizes represent the relative weights of the effects and error bars represent the 95% CI of the regression trend. See the online article for the color version of this figure.

subjects with small numbers of reported effects into a single category. In total, the effects were aggregated into 18 categories (see Table 2). The moderator analysis found no significant moderating role of Subject, $Q(17) = 13.642$, $p = .692$, indicating that the testing effect generalizes across different subjects to an approximately similar degree.

Exam content type. The effects were coded into four subsets by final assessment (exam) content type: Fact learning (e.g., vocabulary words, historical events), Concept learning (e.g., statistical/mathematical concepts, text comprehension), Problem-solving (e.g., how to implement learned clinical techniques in new scenarios), Mixed (i.e., a mixture of different types of learning), and Unknown (not reported). Exam content did not significantly modulate the testing effect, $Q(4) = 8.125$, $p = .087$. Importantly, testing significantly potentiated all types of learning ($g = 0.524$ [0.453, 0.595], $p < .001$, for Fact learning; $g = 0.644$ [0.507, 0.781], $p < .001$, for Concept learning; $g = 0.453$ [0.309, 0.596], $p < .001$, for Problem-solving; $g = 0.424$ [0.285, 0.563], $p < .001$, for Mixed learning), indicating that class quizzing not only facilitates retention of factual knowledge but also boosts conceptual learning and problem-solving. These results run counter to the claim that testing is a purely “drill-and-kill” strategy.

Administration location. Effects were coded into three categories according to whether quizzes were administered in or out of the classroom: In the classroom, Outside the classroom, and Unknown (not reported). Administration location significantly altered the benefits of quizzing, $Q(2) = 10.103$, $p = .006$. Importantly, as shown in Table 2, quizzes administered in the classroom significantly boosted attainment, $g = 0.514$ [0.452, 0.576], $p < .001$,

mitigating the worry that administering quizzes in the classroom may impair learning because it reduces available teaching time.

Next, we explore which type of quiz is more efficient: quizzes administered in versus outside the classroom. After removing the Unknown effects, a further multilevel random-effects meta-regression analysis was conducted, which showed that quizzes administered in the classroom ($g = 0.514$) were associated (at a marginally significant level) with larger learning benefits than those completed outside the classroom ($g = 0.401$ [0.290, 0.512], $p < .001$, $Q(1) = 3.812$, $p = .051$).

Administration timepoint. Based on the timing of quiz administration in relation to the presentation of class content, the effects were coded into three groups: Preclass (quizzes administered before teaching), Postclass (after teaching), and Combination (quizzes administered both before and after teaching). Administration timepoint had a significant influence on the magnitude of the classroom testing effect, $Q(2) = 19.554$, $p < .001$. Importantly, the enhancing effect of preclass quizzes was significant (although modest), $g = 0.186$ [0.036, 0.337], $p = .015$, which goes some way toward settling the debate about whether preclass quizzes can aid student learning.

To explore whether quizzes should be administered before or after the teaching session, a further moderator analysis was conducted, in which Combination effects ($k = 52$) were removed. The results showed that Postclass quizzes ($g = 0.536$ [0.476, 0.597], $p < .001$) induced greater learning gains than Preclass ones ($g = 0.186$), $Q(1) = 8.397$, $p = .004$, indicating that quizzes administered after teaching are more effective than ones presented prior to teaching.

Administration mode. Administration mode was coded into seven categories: Clicker response system, Computer,⁸ Online, Simulation (quizzes simulating a real situation),⁹ Pen-and-paper, Oral (teachers orally deliver quiz questions and students respond orally), and Unknown (not reported). Quiz administration modality did not significantly modulate the testing effect, $Q(6) = 4.297$, $p = .637$, indicating that test-enhanced learning does not depend on administration modality.

Single versus multiple classes. As discussed above, effects obtained from a single class (i.e., a single study-test cycle) might largely result from the direct benefit of testing (i.e., consolidation of studied/tested information), whereas effects obtained from multiple classes (i.e., multiple study-test cycles) might originate from a combination of direct and indirect forward benefits of testing. Hence, it is reasonable to expect larger testing benefits if the testing treatment is implemented across multiple classes than if it is conducted in a single class. To test this expectation, we coded

⁸ Computer quizzes (e.g., Raupach et al., 2016) were typically administered on personal computers at a specific time (e.g., during a class) and place (e.g., in the classroom) and were different from Online ones, which were administered via the internet and students accessed them whenever or wherever they chose.

⁹ For instance, Kromann, Jensen, and Ringsted (2009) assessed the testing effect on medical students' learning of resuscitation skills. In the quizzes, students were provided six cardiac arrest scenarios and required to perform the resuscitation treatment.

the effects into two categories: (a) Single class¹⁰ and (b) Multiple classes, and conducted a multilevel random-effects metaregression analysis. The results confirmed this expectation, with larger learning gains for Multiple classes ($g = 0.541$ [0.476, 0.606], $p < .001$) than Single classes ($g = 0.389$ [0.287, 0.490], $p < .001$, $Q(1) = 6.728$, $p = .010$).

Treatment duration. Recall that the motivation theory makes two contrasting predictions regarding the relationship between treatment duration and test-enhanced learning: (a) a linear increase and (b) an inverted U-shape. Several regression analyses were run to test these two predictions.

Treatment duration was coded into four levels: Single class (treatment administered in a single class session), < Semester (treatment lasting longer than a single class and less than a whole semester), = Semester (treatment lasting a whole semester), and > Semester (treatment lasting longer than a semester). Treatment duration had a significant effect, $Q(3) = 9.174$, $p = .027$. Figure 10 depicts the trend of the classroom testing effect as a function of treatment duration. As clearly shown, testing benefits systematically increased from single class treatments ($g = 0.385$ [0.283, 0.488], $p < .001$), treatments lasting less than a semester ($g = 0.521$ [0.412, 0.631], $p < .001$), treatments lasting a whole semester ($g = 0.547$ [0.465, 0.629], $p < .001$), to those lasting longer than a semester ($g = 0.624$ [0.500, 0.748], $p < .001$). Overall, the results demonstrate a linear increasing function.

Continuous multilevel metaregression analyses were conducted to explore the relationship between the number of class meetings (NCMs; i.e., how many times students met for class) and g s. In total, 338 effects were included in the analyses below, and the remaining 235 effects were excluded because of not reporting NCMs. A linear continuous multilevel random-effects metaregression analysis was conducted to regress g s onto NCMs. The results

demonstrated a positive relationship between them, $\beta = 0.008$ [0.002, 0.014], $Q(1) = 5.899$, $p = .015$ (see Table 3 and Figure 11A), indicating that every additional class meeting increased g by 0.008. This linear increase trend supports the first prediction of the motivation theory.

A quadratic continuous multilevel random-effects metaregression was conducted to assess the second prediction of the motivation theory, in which we regressed g s onto NCMs and the second power of NCMs (i.e., NCM^2). The result showed a U-shaped (rather than inverted U-shaped) function relating NCMs to the classroom testing effect, $Q(2) = 6.860$, $p = .032$ (see Table 3 and Figure 11B), which does not support the second prediction of the motivation theory.

A likelihood ratio (LR) test was run to compare the goodness of fit between the linear and quadratic models and found no significant difference in their fit goodness, $LR = 1.571$, $p = .210$. Because the linear model is simpler than the quadratic model and they did not differ significantly in their fit, we hence conclude that the magnitude of the classroom testing effect increases approximately linearly as a function of NCMs.

Stake. Stake level (i.e., whether quiz performance was incorporated into final course grades or additional awards were offered for superior quiz performance) was coded into three categories: High (performance was incorporated or extra awards were offered), Low (performance was not incorporated nor rewarded), and Unknown (not reported). There was no reliable difference among these three categories, $Q(2) = 3.233$, $p = .199$. Both high-stake ($g = 0.441$ [0.350, 0.531], $p < .001$) and low-stake ($g = 0.477$ [0.380, 0.574], $p < .001$) quizzes significantly aided student learning, and there was no significant difference in their effectiveness, $Q(1) = 0.541$, $p = .462$, when the Unknown category was omitted. Overall, these results imply that stake level plays little role in modulating the testing effect and that even low-stake quizzes can reliably promote learning.

Collaboration. The effectiveness of collaborative and independent quizzing was compared, but no significant difference was detected ($g = 0.653$ [0.472, 0.835], $p < .001$, for collaborative quizzing; $g = 0.490$ [0.432, 0.548], $p < .001$, for independent quizzing), $Q(1) = 3.068$, $p = .080$, implying that independent and collaborative quizzes produce comparable benefits. For reasons to be elucidated in the Discussion, we suggest these results should be interpreted with caution.

Experimental design. Experimental design had two levels: (a) Within-subjects and (b) Between-subjects. Contrary to what both Rowland (2014) (that Between-subjects manipulations are associated with greater learning enhancement than Within-subjects manipulations) and Adesope et al. (2017) (no difference) found, the current analysis found that Within-subjects designs ($g = 0.674$ [0.592, 0.757], $p < .001$) were associated with larger effect sizes than Between-subjects designs ($g = 0.415$ [0.348, 0.481], $p < .001$).

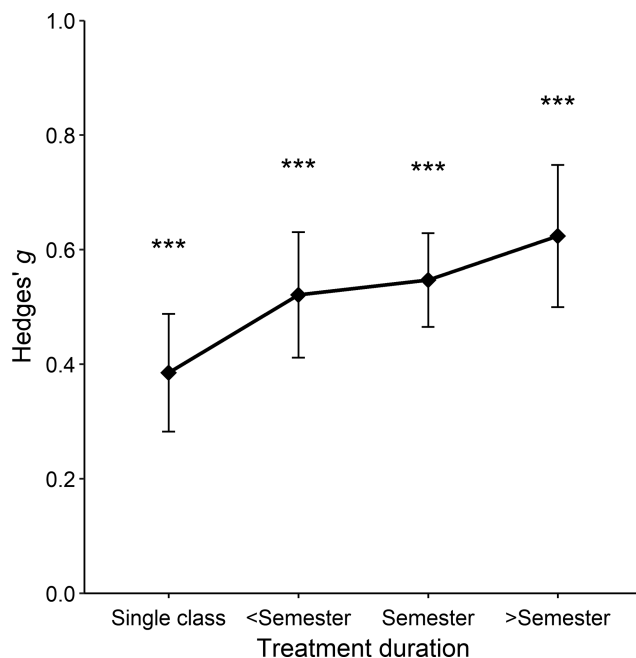


Figure 10. The trend of the testing effect as a function of treatment duration. Error bars represent 95% CI of the weighted mean. *** $p < .001$.

¹⁰ In some studies, researchers (or teachers) might ask students to study some foreign words in a class. Then, in the next class, some students might take a quiz on those studied words while others restudied the words. In a class one week later, all students took a final exam on the words. Even though the study, quiz, and exam phases took place in different classes, such studies were coded as Single class because all words were studied in a single class rather than in repeated study-quiz cycles with different materials studied in each cycle.

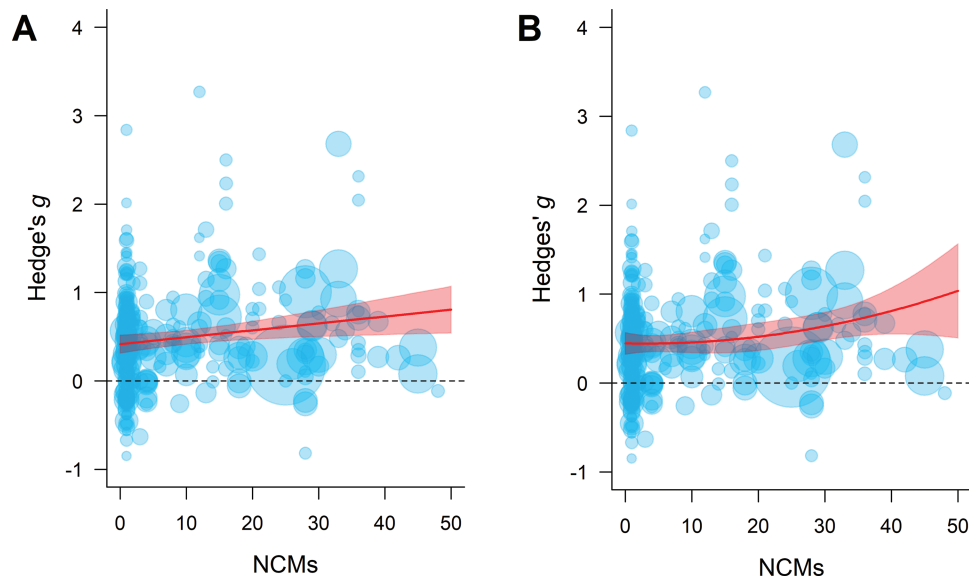


Figure 11. (A) Bubble plot depicting the linear relationship between the testing effect and NCMs (number of class meetings). (B) Quadratic relationship between the testing effect and NCMs. Bubble sizes represent the relative weights of the effects, and error bars represent the 95% CI of the regression trend. See the online article for the color version of this figure.

.001), $Q(1) = 36.037$, $p < .001$. This pattern is consistent with the widely established advantage of within-subjects designs (Gravetter & Forzano, 2011). In such studies, individual differences in students' overall learning abilities and many other psychological characteristics (such as motivation and learning interest) are controlled, minimizing error variance and increasing the estimated effect sizes (Allen, 2017; Thompson & Campbell, 2004).

Instructor consistency. According to whether students in the treatment and control conditions were taught by the same instructor(s), the effects were coded into three categories: Same, Different, and Unknown (not reported). Although the effects were heterogeneous across the three categories, $Q(2) = 9.813$, $p = .007$, there was no significant difference between Same ($g = 0.459$ [0.396, 0.523], $p < .001$) and Different ($g = 0.527$ [0.308, 0.747], $p < .001$) instructors, $Q(1) = 0.307$, $p = .580$, implying little modulating effect of instructor consistency.

Discussion

Test-enhanced learning has been explored in over a century of research and, especially in the past decade, has received an exponential increase of attention. The power of testing in promoting learning is substantial, despite the fact that learners, teachers, and policymakers tend to lack comprehensive metacognitive insight into the beneficial effects of testing. Indeed it is even argued by some that testing (quizzing) in the classroom should be kept to a minimum. Hence, comprehensive reviews are required to provide the best possible evidence about and to promote the practical application of testing. At least 90 reviews have been conducted to pursue these aims. However, only seven of them reported formal meta-analyses. Five mainly focused on laboratory research, which lacks ecological validity and might have limited applicability. For the other two meta-analyses concentrating on the classroom testing

effect, one was published nearly 30 years ago (Bangert-Drowns et al., 1991) and the other only included studies undertaken in high school or university/college Psychology classes (Schwieren et al., 2017). Extending prior research, the current review conducted an up-to-date meta-analysis particularly focusing on the classroom testing effect across the full range of academic subjects and different levels of education.

In total, the current review assessed 19 important questions about test-enhanced learning in the classroom (see Table 1). The findings bear significant implications for practical application of test-enhanced learning and for deepening our understanding of the mechanisms underlying the effect. Below, we first briefly summarize the main research findings and comment on the corresponding practical implications, then discuss theoretical issues, and finally offer some suggestions for future research.

Summary of Research Findings and Practical Implications

The meta-analysis found that testing (class quizzing) reliably enhances academic performance on criterion achievement assessments, and assessment scores are raised to a medium extent ($g = 0.499$). A positive effect of testing was detected in a majority (82.9%) of the included studies. The p -curve analysis demonstrated strong evidence supporting the existence of the overall effect. Five publication bias assessment tools (publication year, publication status, PET-PEESE, trim-and-fill, and 3PSM) were employed to detect potential bias in the included effects, and none revealed noteworthy evidence of publication bias.

The magnitude of the meta-analytic effect ($g = 0.499$) merits some commentary. Some recent large sample studies demonstrated that the SD s of grade point average (GPA) in high schools and universities are about 0.60–0.70 (e.g., Fajnzylber, Lara, & León,

2019; Westrick, 2017). Hence, an effect size of $g = 0.499$ can be roughly translated into a change in GPA of about 0.30–0.35 points. This is equivalent to moving from the 50th to the 69th percentile of a normal distribution. Although $g = 0.499$ is only a medium-sized effect according to conventional descriptors, it is a notably large effect by the standards of educational interventions. An analysis of more than 800 educational meta-analyses found an average effect size of 0.40 (Hattie, 2009), and a meta-analysis of 141 high-quality randomized control trials commissioned by two major education funding agencies found a mean effect size of 0.06, with only one trial obtaining an effect size larger than 0.5 (Lortie-Forgues & Inglis, 2019). In addition, as noted by many educators and researchers, “in real-world [educational] settings, a fifth of a standard deviation [0.20 *SD*] is a large effect” (Dynarski, 2017; Yeager et al., 2019). Against these comparators, the relatively large effect of testing on classroom learning is noteworthy and of considerable practical importance.

Categorical and continuous metaregression analyses were conducted to explore the moderators and boundary conditions of test-enhanced learning in the classroom. The results showed that testing was either significantly or numerically more effective for enhancing learning than many other strategies, such as no/filler activity ($g = 0.610$), testing with fewer questions ($g = 0.465$), restudying ($g = 0.330$), and other elaborative strategies ($g = 0.095$). Although the effects were modest when comparing testing with other elaborative strategies (which were relatively few in number), the enhancement effect demonstrates that testing is numerically superior for promoting knowledge acquisition and retention (Karpicke & Blunt, 2011; Lechuga, Ortega-Tudela, & Gómez-Ariza, 2015).

Test format in quizzes significantly modulated the magnitude of test-enhanced learning, with matching tests producing the largest benefit and free recall tests the smallest. No significant difference in effectiveness was detected between recall and recognition tests, regardless of whether feedback was provided (recall: $g = 0.541$; recognition: $g = 0.607$) or not (recall: $g = 0.520$; recognition: $g = 0.518$), implying that test difficulty plays little role in modulating the classroom testing effect. Test format matching between quizzes and final assessments significantly modulated the magnitude of the testing effect, with larger enhancement for consistent ($g = 0.531$) than for inconsistent formats ($g = 0.399$), suggesting that quizzing tends to be more beneficial when test formats between quizzes and exams are matched.

Quizzing not only benefitted memorization of tested ($g = 0.512$) but also of untested materials ($g = 0.321$), indicating the transferability of the testing effect to untested materials. The significant enhancing effect on untested materials is inconsistent with the null effect reported by Pan and Rickard (2018). As discussed in the Introduction, two explanations may explain this divergence. The first is that the current review included a greater number of effects ($k = 43$) than in Pan and Rickard (2018) ($k = 17$), which means that the statistical power should be larger to detect a significant enhancing effect in the current review. The second explanation is semantic coherence, which has been shown to be an essential requirement to observe an enhancing effect on untested material (Little et al., 2011). Different sections of lecturing contents (and textbooks) are deeply related, but the stimuli in many of Pan and Rickard’s (2018) studies were unrelated. Besides these two explanations, another potential source for this divergence is the differ-

ence in experimental procedures between laboratory and classroom research.

As discussed in the Introduction, classroom studies were typically conducted across multiple classes and involved multiple study-test cycles with different materials in each cycle, whereas laboratory studies generally involved only a single study-test cycle. Hence, besides the classic benefits of testing (i.e., testing consolidates studied information), the indirect forward benefits of testing (i.e., testing boosts new learning) may also contribute importantly to the classroom testing effect. The forward enhancing effect of testing can help to account for the divergent findings detected by the current meta-analysis and that of Pan and Rickard (2018). For instance, in classroom research, testing prospectively facilitated learning of new information in subsequent classes, and hence new materials studied in subsequent classes could benefit from prior quizzes regardless of whether they were tested or not in the subsequent quizzes.

Offering corrective feedback ($g = 0.537$) increased learning gains compared with not providing feedback ($g = 0.374$), encouraging the recommendation that corrective feedback should routinely be provided following quizzing. Categorical and continuous metaregression analyses jointly demonstrated that the more occasions on which class content is tested, the larger the learning gains, suggesting that learners and instructors can profitably employ repeated retrieval practice. Although this finding is inconsistent with what Adesope et al. (2017) found, it is closely in line with the empirical findings documented by dozens of studies, showing that the magnitude of test-induced enhancement increases with the amount of repeated retrieval (e.g., Eriksson et al., 2011; Karpicke & Roediger, 2007b, 2008; Kornell & Bjork, 2008; McDermott, 2006; Pyc & Rawson, 2009; Roediger & Karpicke, 2006b; Soderstrom et al., 2016; Vaughn & Rawson, 2011; Wheeler & Roediger, 1992).

Testing produced significant learning gains in elementary school ($g = 0.328$), middle school ($g = 0.597$), high school ($g = 0.655$), and university/college ($g = 0.486$) classes, but no significant enhancing effect was detected for continuing education ones ($g = 0.314 [-0.081, 0.709]$, $p = .119$). This nonsignificant effect might result from the small number of included effects ($k = 6$). In addition, all the effects for continuing education originated from a single class (workshop) treatment. As demonstrated above, the magnitude of test-enhanced learning systemically increases as a function of treatment duration. Hence, the modest enhancement for single class treatments might be another source of the modest testing effect in continuing education classes.

Although the modulating roles of gender in a variety of memory phenomena have been established and its role in test-enhanced learning has also been repeatedly explored (but with inconsistent findings), no meta-analyses have investigated this variable. To fill this gap, we regressed 211 effects onto their corresponding female gender ratios and found a nonsignificant relationship, implying equivalent benefits of testing for male and female students.

Combining the effects from 31 subjects into 18 academic subject categories, we observed that test-enhanced learning did not significantly vary across subjects, implying strong generalizability of the testing effect. A regression analysis was conducted to investigate the claim that quizzing is a “drill-and-kill” strategy which only enhances “inert knowledge” (Fey, 2012). The results showed that testing was not only beneficial for learning factual

knowledge ($g = 0.524$) but also promoted conceptual learning ($g = 0.644$) and facilitated problem-solving ($g = 0.453$). Hence, these results counter the view that testing only consolidates inert knowledge, it can also facilitate comprehension (i.e., knowledge organization and integration; e.g., Jing et al., 2016; Roediger & Karpicke, 2006a) and knowledge transfer to aid solving new problems in unfamiliar contexts (e.g., Butler, 2010; Jacoby et al., 2010; Lee & Ahn, 2018; Yang & Shanks, 2018).

Further meta-analyses were conducted to explore where (location), when (timepoint), and how (mode) to administer quizzes to optimize the beneficial effects of testing. The results showed that quizzes administered in the classroom ($g = 0.514$) tend to be more effective than those administered outside the classroom ($g = 0.401$), implying that classroom quizzes supervised by instructors are more beneficial than those administered outside the classroom without direct supervision. Quizzes administered after teaching (postclass: $g = 0.536$) were more effective than those presented prior to teaching (preclass: $g = 0.186$). Administration mode had little impact, implying that the beneficial effects of testing are not tied to the mode through which retrieval practice is administered but rather to retrieval practice itself.

It is worth noting that the significant enhancing effect of testing inside the classroom ($g = 0.514$) tends to mitigate the concern that class quizzing displaces other class activities and impairs attainment. Stated differently, even though it should be acknowledged that administering a class quiz “borrows” time from didactic teaching, the meta-analysis suggests that this tradeoff between teaching and quizzing seems to be worthwhile.

As hypothesized, effects obtained from multiple class treatments (i.e., involving repeated study-test cycles, with new materials studied in each cycle; $g = 0.541$) were significantly larger than those derived from a single class treatment ($g = 0.389$). More importantly, we also observed that quizzing benefits increased from single class treatments ($g = 0.385$), treatments lasting less than a semester ($g = 0.521$), treatments lasting a whole semester ($g = 0.547$), to those lasting greater than a semester ($g = 0.624$). This linear increase was confirmed by a continuous regression analysis, which regressed g s onto the total number of class meetings. These consistent findings suggest that the longer the testing treatment, the larger the learning gain.

Both low-stake ($g = 0.477$) and high-stake ($g = 0.441$) quizzes significantly promoted academic achievement, with little difference in their effectiveness. Regardless of whether students answered quiz questions independently ($g = 0.490$) or collaboratively with their peers ($g = 0.653$), testing was always beneficial and there was no significant difference between independent and collaborative testing. The difference between within-subjects ($g = 0.674$) and between-subjects ($g = 0.415$) designs was significant. Instructor consistency between the treatment and control conditions did not significantly modulate the testing effect (matched instructors: $g = 0.459$; different instructors: $g = 0.527$).

Overall, the above-summarized findings demonstrate the power of testing to promote students’ learning in the classroom. Although testing has been met with occasional skepticism and criticism, these objections do not outweigh the extensive merits of frequent testing.

As discussed above, metacognitive unawareness of test-enhanced learning, anxiety about the tradeoff between didactic teaching and class quizzing, and concerns that testing may only

benefit inert knowledge might partially explain underemployment of test-enhanced learning in the classroom. Besides these factors, there may be others. For instance, some instructors may assume that students with low learning ability will benefit less or even suffer from class quizzing as their quiz performance is typically low, leading to insufficient exposure to class materials. This concern may constitute a further worry that class quizzing would exacerbate individual differences in academic performance among students with different levels of learning ability.

Even though the modulating roles of many individual differences constructs in test-enhanced learning have not been assessed in the current meta-analysis, many other studies have provided evidence to mitigate those concerns (e.g., Brewer & Unsworth, 2012; Pan, Pashler, Potter, & Rickard, 2015; Yang et al., 2020). For instance, Yang et al. (2020) recently conducted a large sample (1,032 participants) study to investigate the modulating effect of working memory capacity (WMC) on test-enhanced learning, and the results showed that individuals with low WMC benefit more from retrieval practice than those with high WMC (for related findings, see Agarwal, Finley, Rose, & Roediger, 2017), implying that testing can narrow, rather than exaggerate, individual differences in learning efficiency based on WMC. There are a few other studies demonstrating that individuals with different levels of WMC benefit equally from retrieval practice (e.g., Brewer & Unsworth, 2012; Wiklund-Hörnqvist, Jonsson, & Nyberg, 2014). Brewer and Unsworth (2012) revealed that individuals with low episodic memory ability benefit more from testing than those with high ability, and Pan et al. (2015) found that the testing effect generalizes across a range of episodic memory abilities. Brewer and Unsworth (2012) observed a larger testing effect for individuals with lower general-fluid intelligence. In total, these findings jointly underline that individuals with inferior learning or cognitive ability benefit equally or even more from retrieval practice compared with those with high ability.

Another concern instructors may have is that frequent testing provokes test anxiety (D. Steele, 2011), which is a major cause of learning difficulty (Hembree, 1988). However, recent findings run counter to this concern. For instance, Yang et al. (2020) showed strong evidence that low-stake tests have minimal influence on test anxiety. Szpunar et al. (2013) observed that frequent tests significantly allay test anxiety. A few large-scale surveys have demonstrated that the majority of college (e.g., over 90% in Sullivan, 2017) and middle school students (e.g., 70% in Agarwal et al., 2014) believe that class quizzes are likely to reduce their test anxiety.

In summary, some practitioners may have concerns about the negative influences of testing, leading to test-enhanced learning not being applied as widely as it could be. However, the current review and other recent studies jointly provide evidence to mitigate these concerns. Hence, practitioners are encouraged to consider testing as a learning tool, instead of simply regarding it as an assessment technique.

Theoretical Implications

Theoretical analysis of test-enhanced learning has largely been based on laboratory research, while the majority of classroom studies have focused instead on its practical applications. Hence a novel contribution of the current review is that it aimed to test four

major theoretical accounts of the testing effect against data collected in the field: (a) additional exposure, (b), transfer-appropriate processing, (c) retrieval effort, and (d) motivation.

Additional exposure. The additional exposure account hypothesizes that testing boosts learning and retention via providing greater reexposure in the treatment condition. Its major prediction has been corroborated: Testing was associated with substantially larger enhancement when compared with no/filler activity, with low reexposure ($g = 0.610$), than when compared with restudying, with high reexposure ($g = 0.330$). Moreover, another line of evidence also strongly supports this account: quizzing followed by corrective feedback (high reexposure; $g = 0.537$) generated greater benefits than quizzing without corrective feedback (low reexposure; $g = 0.374$). Although both lines of evidence provide strong support for the additional exposure account, it must be acknowledged that it cannot be the only explanation because quizzing also more effectively boosts learning attainment compared with restudying ($g = 0.330$).

Transfer-appropriate processing. The transfer-appropriate processing explanation assumes that testing facilitates subsequent retrieval because the contexts in the initial learning and final assessment phases are more similar in the treatment condition than in the control condition. Accordingly, it predicts that the magnitude of testing benefits ought to be larger when the test formats in the acquisition and final assessment phases are matched. This prediction was confirmed by the finding that consistent test formats ($g = 0.531$) were associated with a significantly larger effect size than mismatched formats ($g = 0.399$).

Retrieval effort. The retrieval effort theory proposes that the level of test-enhanced learning is dependent on how demanding the retrieval processes are in the initial tests. It predicts that difficult tests (e.g., recall tests) should more effectively boost memory storage and retrieval strength than easy tests (e.g., recognition tests). However, the current analysis observed that recognition tests ($g = 0.518$) produced equivalent enhancement as recall tests ($g = 0.520$). This finding must be interpreted with caution as additional exposure might contribute to the null difference: Some tests provided no feedback, hence reducing exposure to the learning materials. To mitigate the possibility that feedback provision was confounded with test type, another analysis was conducted on the effects for which corrective feedback was offered (hence equating exposure). The result again showed little difference between recall ($g = 0.541$) and recognition ($g = 0.607$) tests (and if anything, in the wrong direction), failing to support the retrieval effort explanation (for related findings, see Adesope et al., 2017).

This approximate equivalence of the benefits of recognition and recall tests is important for a practical reason: There are several merits of recognition tests in the classroom. Answering recognition questions is quicker and hence such tests save time for teaching, and recognition quizzes are easier to administer and score with the help of technology. For instance, multiple-choice tests are frequently administered using clicker response systems or smartphones. Immediately following each question, such systems can automatically score and summarize students' responses, which then act as diagnostic feedback guiding teachers to provide corrective feedback and deliver a further illustration of the content that students do not master well. In summary, the above results do not support the retrieval effort theory's explanation of the classroom testing effect. However, it merits further examination.

Motivation. The motivation theory provides a viable account of the indirect forward benefits of testing (Yang et al., 2018): Testing stimulates learners to commit more effort in the subsequent learning process, which in turn boosts new learning. This theory has been tested by three analyses in the current work. The first analysis found that effects obtained across multiple classes ($g = 0.541$) were larger than those derived from a single class ($g = 0.389$). The second is the observation that the learning gain significantly increased from single class treatments to those lasting longer than a semester. This is reconfirmed by the third analysis, which showed a positive relationship between the testing effect and NCMs. Overall, these findings supply consistent evidence supporting the motivation theory as an account for test-enhanced learning in the classroom.

A categorical regression analysis showed little difference between high ($g = 0.441$) and low ($g = 0.477$) stake quizzes. A tempting inference from this finding is that it violates the motivation explanation as high-stake tests did not produce greater learning gains than low-stake ones. We highlight however that the null difference does not directly counter the motivation theory. As discussed in the Introduction, high-stake quizzes may provoke test anxiety, which in turn counters the merits of higher motivation, leading to a null difference between high- and low-stake quizzes (Khanna, 2015). Supporting this explanation, Khanna (2015) observed that, in her Introductory Psychology course, students taking ungraded (low-stake) quizzes excelled in the final course exam compared with those taking graded (high-stake) quizzes and those taking no quizzes.

The distinction between intrinsic and extrinsic motivation should also be taken into account when interpreting the null modulating role of test stake (Ryan & Deci, 2000). Increasing test stake may raise extrinsic motivation. It is well known that the consequences of extrinsic motivation tend to be fickle and students can quickly lose interest in external rewards (Blake, 2015). Hence, increasing extrinsic motivation (by raising test stake) may have little impact on long-term learning outcomes in the classroom. Instead of affecting learning by boosting extrinsic motivation, it is possible that testing may increase students' intrinsic motivation to seek knowledge for its own sake. For instance, dissatisfaction about retrieval failures in prior quizzes and/or realization of the gap between actual learning progress and desired status may intrinsically drive students to study.

Although specific motivational mechanisms of the testing effect have been developed and investigated, this work has made remarkably little connection with the broader and very influential literature on motivation to learn (Schunk, Meece, & Pintrich, 2012). For example, it is plausible to conjecture that a practice test might influence both expectation of success in a later assessment as well as its perceived value, the two key components of the expectancy-value theory of motivation (Wigfield & Eccles, 2000). Equally, practice tests might boost self-efficacy (Bandura, 1986; Schunk, 1991). Yet these potential links have rarely been considered or developed in the studies included in the meta-analysis. Conversely, although these general theories of motivation have inspired many interventions, little attention has been paid to testing as a means of boosting educational motivation (e.g., Lazowski & Hulleman, 2016).

In summary, the additional exposure, transfer-appropriate processing, and motivation theories are viable accounts of the class-

room testing effect. Although no supporting evidence was obtained for the retrieval effort theory, we reiterate that the current findings do not conclusively refute it.

Future Research Directions

Although test-enhanced learning has been investigated in hundreds of studies over the last century, there remain at least five major questions awaiting further investigation. First, continuing education only contributed six effects to the current meta-analysis, suggesting that more work is needed to explore the testing effect on professional learning.

Second, an interesting outcome of the meta-analysis is the finding that testing was only marginally better than other elaborative strategies ($g = 0.095$, $p = .062$). Of course, no proponent of testing in the classroom would claim that it is the only method for enhancing learning and retention. Elaborative strategies included many methods, such as concept mapping, note-taking, summarizing, and so on, and there were relatively few effect sizes for any of these individual methods. Future research to assess in more detail what strategies match testing in effectiveness would be valuable (Heitmann, Grund, Berthold, Fries, & Roelle, 2018; Karpicke & Blunt, 2011; Rummer et al., 2017). An intriguing question is whether additive effects can be achieved by combining testing with one or another elaborative strategy (Karpicke & Bauernschmidt, 2011; Karpicke & Roediger, 2007a).

Third, although the testing effect has been extensively explored in several subjects (e.g., $k = 144$ for Psychology and $k = 39$ for Medical Science), some subjects have received little attention (e.g., $k = 3$ for Law, $k = 2$ for Computer Science, and $k = 1$ for Physics), and to our knowledge its effectiveness has never been evaluated in many other subjects (e.g., Philosophy, Agriculture, and Archival Science). Hence, the generality of test-enhanced learning to other subject domains awaits future exploration.

Fourth, even though the current review detected no significant difference between collaborative and independent quizzing, the results did show a clear trend that collaborative quizzing ($g = 0.653$) tends to be more effective than independent quizzing ($g = 0.490$). The nonsignificant difference might result from the fact that far fewer studies ($k = 21$) employed collaborative than independent quizzes ($k = 552$). Future research could profitably pay more attention to collaborative testing and the possibility that its social dimension amplifies the benefits of retrieval practice.

Lastly, but importantly, future research should consider how to translate principles from cognitive research on test-enhanced learning into mainstream educational policy and practice. The “know–do gap,” a well-known phenomenon referring to the gap between what we learn from research and what is applied in the field (Bennett & Jessani, 2011), is highly relevant to the topic of test-enhanced learning. For instance, testing is frequently recognized as an assessment of learning rather than an assessment for learning, and test-enhanced learning has not been practically implemented as widely as it could be (Geller et al., 2018; Hartwig & Dunlosky, 2012; Kornell & Bjork, 2007; McAndrew et al., 2016; Morehead et al., 2016; Yan et al., 2014). How to bridge the existing gap between research on test-enhanced learning and its practical application is a challenge that has not been systematically studied. A priority to narrow this gap should be to foster commu-

nication among researchers, learners, instructors, and policymakers (Bennett & Jessani, 2011). For instance, researchers can endeavor to translate their research into short and accessible articles or videos to make it easier to understand and more widely available. Equally if not more important is to develop and evaluate interventions to boost the employment of quizzing in the classroom.

Concluding Remarks

The current review utilized multilevel random-effects models to quantitatively synthesize 573 effects from 222 research reports comprising data from 48,478 students to evaluate test-enhanced learning in the classroom. The research findings shed light on practical applications and theoretical explanations of the testing effect. In brief, the takeaway messages are:

1. Testing, by comparison with other strategies (such as no/filler activity, restudying, and other elaborative strategies), overall boosts student attainment, and the enhancement is to a medium-sized extent ($g = 0.499$).
2. Test-enhanced learning generalizes to a variety of test formats.
3. In the classroom, testing not only consolidates retention of content that is directly tested but also promotes memorization of untested knowledge.
4. Presenting corrective feedback following quizzing enhances the mnemonic benefits of testing.
5. The more occasions class content is tested, the larger the learning gains.
6. The testing effect occurs in primary (elementary), secondary (middle and high), and postsecondary (university/college) education. The effect in continuing education requires more research.
7. Male and female students obtain comparable learning benefits from testing.
8. The enhancing effect of testing applies across 18 academic subject categories.
9. Testing not only enhances learning of facts but also facilitates knowledge comprehension and application.
10. Quizzes administered in the classroom tend to be more beneficial than ones administered outside the classroom. Quizzes administered after teaching more effectively boost attainment than ones administered prior to teaching. Administration mode has minimal influence on the magnitude of the testing effect.
11. The longer the testing treatment, the larger the learning gains.
12. Stake level plays little moderating role in the classroom testing effect.

13. At the current stage, no firm conclusions can be drawn regarding the superiority of independent or collaborative testing, and further exploration will be useful.
14. Testing treatments manipulated within-subjects are associated with larger effect sizes than ones manipulated between-subjects.
15. Instructor consistency between the treatment and control conditions does not significantly affect the magnitude of the testing effect.
16. The additional exposure, transfer-appropriate processing, and motivation theories are viable accounts of the classroom testing effect. The retrieval effort account receives less support.

References

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

- Abbott, E. E. (1909). On the analysis of the factor of recall in the learning process. *The Psychological Review: Monograph Supplements*, 11, 159–177. <http://dx.doi.org/10.1037/h0093018>
- *Achord, R. L. K. (2015). *The effect of frequent quizzing on student populations with differing preparation and motivation in the high school biology classroom* (Master's thesis). Retrieved from https://digitalcommons.lsu.edu/gradschool_theses/871
- Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the use of tests: A meta-analysis of practice testing. *Review of Educational Research*, 87, 659–701. <http://dx.doi.org/10.3102/0034654316689306>
- *Agarwal, P. K. (2011). *Examining the relationship between fact learning and higher order learning via retrieval practice* (Doctoral dissertation). Washington University, St. Louis, MO. Retrieved from <https://openscholarship.wustl.edu/etd/546/>
- Agarwal, P. K., D'Antonio, L., Roediger, H. L., III, McDermott, K. B., & McDaniel, M. A. (2014). Classroom-based programs of retrieval practice reduce middle school and high school students' test anxiety. *Journal of Applied Research in Memory & Cognition*, 3, 131–139. <http://dx.doi.org/10.1016/j.jarmac.2014.07.002>
- Agarwal, P. K., Finley, J. R., Rose, N. S., & Roediger, H. L., III. (2017). Benefits from retrieval practice are greater for students with lower working memory capacity. *Memory*, 25, 764–771. <http://dx.doi.org/10.1080/09658211.2016.1220579>
- Agarwal, P. K., & Roediger, H. L., III. (2011). Expectancy of an open-book test decreases performance on a delayed closed-book test. *Memory*, 19, 836–852. <http://dx.doi.org/10.1080/09658211.2011.613840>
- *Agarwal, P. K., Roediger, H. L., McDaniel, M. A., & McDermott, K. B. (2008, May). *Improving learning with classroom quizzes*. Paper presented at the 20th Annual Meeting of the Association for Psychological Science. Retrieved from https://www.researchgate.net/publication/237501501_Improving_Learning_With_Classroom_Quizzes_With_Classroom_Quizzes
- *Alade, O. M., & Kuku, O. O. (2017). Impact of frequency of testing on study habits and achievement in mathematics among secondary school students in Ogun State, Nigeria. *Journal of Educational Research and Practice*, 7, 1–18. <http://dx.doi.org/10.5590/JERAP.2017.07.1.01>
- *Alimoradi, M., Mahmoodi, K., & Rajabi, P. (2015). The effect of immediate grammar tests on the improvement of Iranian pre-university students' final exam achievement. *International Journal of Engineering Education*, 4, 269–279.
- Allen, M. (2017). *The SAGE encyclopedia of communication research methods*. London, UK: SAGE Publications. <http://dx.doi.org/10.4135/9781483381411>
- *Alzughairi, M., Alotaibi, M., Ahmed, F., Alqahtani, B., & Bargo, M. (2016). PBL quizzes and their effects on student performance. *Journal of U. S.-China Medical Science*, 13, 108–112. <http://dx.doi.org/10.17265/1548-6648/2016.02.007>
- *Appaji, A. C., & Kulkarni, R. (2012). Multiple choice questions as a teaching learning tool in addition to assessment method. *National Journal of Integrated Research in Medicine*, 3, 91–95.
- *Arteaga, I. L., & Vinken, E. (2013). Example of good practice of a learning environment with a classroom response system in a mechanical engineering bachelor course. *European Journal of Engineering Education*, 38, 652–660. <http://dx.doi.org/10.1080/03043797.2012.719000>
- Aslan, A., & Bäuml, K. H. T. (2016). Testing enhances subsequent learning in older but not in younger elementary school children. *Developmental Science*, 19, 992–998. <http://dx.doi.org/10.1111/desc.12340>
- Asperholm, M., Högman, N., Rafi, J., & Herlitz, A. (2019). What did you do yesterday? A meta-analysis of sex differences in episodic memory. *Psychological Bulletin*, 145, 785–821. <http://dx.doi.org/10.1037/bul0000197>
- *Atia, A., Ashour, A., & Abired, A. (2018). Frequent announced pharmacology quizzes have no impact on academic performance: An exploratory study. *Ibnosina Journal of Medicine and Biomedical Sciences*, 10, 18–20. http://dx.doi.org/10.4103/ijmbs.ijmbs_44_17
- *Avci, G. (2011). *Transfer of the testing effect: Just how powerful is it?* (Doctoral dissertation). Rice University, Houston, TX. Retrieved from <https://hdl.handle.net/1911/64378>
- *Ayyub, A., & Mahboob, U. (2017). Effectiveness of Test-Enhanced Learning (TEL) in lectures for undergraduate medical students. *Pakistan Journal of Medical Sciences*, 33, 1339–1343. <http://dx.doi.org/10.12669/pjms.336.13358>
- *Azorlosa, J. L. (2011). The effect of announced quizzes on exam performance: II. *Journal of Instructional Psychology*, 38, 3–7.
- *Azorlosa, J. L., & Renner, C. H. (2006). The effect of announced quizzes on exam performance. *Journal of Instructional Psychology*, 33, 278–283.
- *Bachman, L., & Bachman, C. (2011). A study of classroom response system clickers: Increasing student engagement and performance in a large undergraduate lecture class on architectural research. *Journal of Interactive Learning Research*, 22, 5–21.
- *Baghdady, M., Carnahan, H., Lam, E. W. N., & Woods, N. N. (2014). Test-enhanced learning and its effect on comprehension and diagnostic accuracy. *Journal of Instructional Psychology*, 33, 278–283. <http://dx.doi.org/10.1111/medu.12302>
- *Balch, W. R. (1998). Practice versus review exams and final exam performance. *Teaching of Psychology*, 25, 181–185. http://dx.doi.org/10.1207/s15328023top2503_3
- Bandura, A. (1986). The explanatory and predictive scope of self-efficacy theory. *Journal of Social and Clinical Psychology*, 4, 359–373. <http://dx.doi.org/10.1521/jsocp.1986.4.3.359>
- Bangert-Drowns, R. L., Kulik, J. A., & Kulik, C.-L. C. (1991). Effects of frequent classroom testing. *The Journal of Educational Research*, 85, 89–99. <http://dx.doi.org/10.1080/00220671.1991.10702818>
- *Barcroft, J. (2007). Effects of opportunities for word retrieval during second language vocabulary learning. *Language Learning*, 57, 35–56. <http://dx.doi.org/10.1111/j.1467-9922.2007.00398.x>
- *Barcroft, J. (2015). Can retrieval opportunities increase vocabulary learning during reading? *Foreign Language Annals*, 48, 236–249. <http://dx.doi.org/10.1111/flan.12139>
- *Batsell, W. R., Jr., Perry, J. L., Hanley, E., & Hostetter, A. B. (2017). Ecological validity of the testing effect: The use of daily quizzes in Introductory Psychology. *Teaching of Psychology*, 44, 18–23. <http://dx.doi.org/10.1177/0098628316677492>

- Bäuml, K.-H. T., & Kliegl, O. (2013). The critical role of retrieval processes in release from proactive interference. *Journal of Memory and Language*, 68, 39–53. <http://dx.doi.org/10.1016/j.jml.2012.07.006>
- *Beckman, W. S. (2007). Pre-testing as a method of conveying learning objectives. *Journal of Aviation/Aerospace Education Research*, 17, 61–70. <http://dx.doi.org/10.15394/jaer.2008.1447>
- *Bego, C. R., Lyle, K. B., Ralston, P. A., & Hieb, J. L. (2017, October). *Retrieval practice and spacing in an engineering mathematics classroom: Do the effects add up?* Paper presented at the 2017 IEEE Frontiers in Education Conference (FIE).
- *Bell, M. C., Simone, P. M., & Whitfield, L. C. (2015). Failure of online quizzing to improve performance in introductory psychology courses. *Scholarship of Teaching and Learning in Psychology*, 1, 163–171. <http://dx.doi.org/10.1037/stl0000020>
- Bennett, G., & Jessani, N. (2011). *The knowledge translation toolkit: Bridging the know-do gap: A resource for researchers*. New Delhi, India: SAGE Publications India. <http://dx.doi.org/10.4135/9789351507765>
- *Betsch, T., Quittenbaum, N., & Lüders, M. (2015). On the robustness of the quizzing effect under real teaching conditions. *Zeitschrift für Pädagogische Psychologie / German Journal of Educational Psychology*, 29, 109–114. <http://dx.doi.org/10.1024/1010-0652/a000149>
- *Bhatt, M., & Dua, S. (2016). Use of multiple choice questions during lectures helps medical students improve their performance in written formative assessment in physiology. *National Journal of Physiology, Pharmacy and Pharmacology*, 6, 576–580.
- *Bhatt, M., Thapa, B., Bhattacharya, A., Bhinganiya, P., Minhas, S., & Sharma, S. (2015). MCQ supplementation in a physiology didactic class: A learning tool. *National Journal of Integrated Research in Medicine*, 6, 72–76.
- *Bing, S. B. (1984). Effects of testing versus review on rote and conceptual learning from prose. *Instructional Science*, 13, 193–198. <http://dx.doi.org/10.1007/BF00052385>
- 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). New York, NY: Worth.
- *Bjork, E. L., Little, J. L., & Storm, B. C. (2014). Multiple-choice testing as a desirable difficulty in the classroom. *Journal of Applied Research in Memory & Cognition*, 3, 165–170. <http://dx.doi.org/10.1016/j.jarmac.2014.03.002>
- Bjork, R. A. (1994). *Learning, remembering, believing: Enhancing human performance*. Washington, DC: National Academies Press.
- Blake, C. (2015). *Cultivating motivation: How to help students love learning*. Retrieved from <https://resilienteducator.com/classroom-resources/cultivating-student-motivation/>
- Blaxton, T. A. (1989). Investigating dissociations among memory measures: Support for a transfer-appropriate processing framework. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15, 657–668. <http://dx.doi.org/10.1037/0278-7393.15.4.657>
- *Bobby, Z., & Meiyappan, K. (2018). “Test-enhanced” focused self-directed learning after the teaching modules in biochemistry. *Biochemistry and Molecular Biology Education*, 46, 472–477. <http://dx.doi.org/10.1002/bmb.21171>
- *Bojinova, E., & Oigara, J. (2011). Teaching and learning with clickers: Are clickers good for students? *Interdisciplinary Journal of E-Learning and Learning Objects*, 7, 169–184. Retrieved from <https://www.learntechlib.org/p/44737/>
- Borenstein, M., & Cooper, H. (2009). *The handbook of research synthesis and meta-analysis* (Vol. 2). New York, NY: Russell Sage Foundation.
- Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). Converting among effect sizes. In U. Chichester (Ed.), *Introduction to meta-analysis* (pp. 45–49). Hoboken, NJ: Wiley. <http://dx.doi.org/10.1002/9780470743386.ch7>
- *Bouwmeester, S., & Verhoeven, 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*, 26, 61–74. <http://dx.doi.org/10.1007/s10212-010-0036-3>
- Brewer, G. A., & Unsworth, N. (2012). Individual differences in the effects of retrieval from long-term memory. *Journal of Memory and Language*, 66, 407–415. <http://dx.doi.org/10.1016/j.jml.2011.12.009>
- *Brojde, C. L., & Wise, B. W. (2008). *An evaluation of the testing effect with third grade students*. Retrieved from <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.547.2878>
- *Brown, G. A., Bice, M. R., Shaw, B. S., & Shaw, I. (2015). Online quizzes promote inconsistent improvements on in-class test performance in introductory anatomy and physiology. *Advances in Physiology Education*, 39, 63–66. <http://dx.doi.org/10.1152/advan.00064.2014>
- *Brown, M. J., & Tallon, J. (2015). The effects of pre-lecture quizzes on test anxiety and performance in a statistics course. *Education*, 135, 346–350.
- Bufe, J., & Aslan, A. (2018). Desirable difficulties in spatial learning: Testing enhances subsequent learning of spatial information. *Frontiers in Psychology*, 9, 1701. <http://dx.doi.org/10.3389/fpsyg.2018.01701>
- *Burdo, J., & O'Dwyer, L. (2015). The effectiveness of concept mapping and retrieval practice as learning strategies in an undergraduate physiology course. *Advances in Physiology Education*, 39, 335–340. <http://dx.doi.org/10.1152/advan.00041.2015>
- Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 36, 1118–1133. <http://dx.doi.org/10.1037/a0019902>
- *Campbell, J., & Mayer, R. E. (2009). Questioning as an instructional method: Does it affect learning from lectures? *Applied Cognitive Psychology*, 23, 747–759. <http://dx.doi.org/10.1002/acp.1513>
- 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. <http://dx.doi.org/10.3758/BF03193405>
- *Carpenter, S. K., Lund, T. J. S., Coffman, C. R., Armstrong, P. I., Lamm, M. H., & Reason, R. D. (2016). A classroom study on the relationship between student achievement and retrieval-enhanced learning. *Educational Psychology Review*, 28, 353–375. <http://dx.doi.org/10.1007/s10648-015-9311-9>
- *Carpenter, S. K., Pashler, H., & Cepeda, N. J. (2009). Using tests to enhance 8th grade students' retention of U.S. history facts. *Applied Cognitive Psychology*, 23, 760–771. <http://dx.doi.org/10.1002/acp.1507>
- *Carpenter, S. K., Rahman, S., & Perkins, K. (2018). The effects of prequestions on classroom learning. *Journal of Experimental Psychology: Applied*, 24, 34–42. <http://dx.doi.org/10.1037/xap0000145>
- *Carrillo-de-la-Peña, M. T., Baillès, E., Caseras, X., Martínez, À., Ortet, G., & Pérez, J. (2009). Formative assessment and academic achievement in pre-graduate students of health sciences. *Advances in Health Sciences Education*, 14, 61–67. <http://dx.doi.org/10.1007/s10459-007-9086-y>
- Carter, E. C., Schönbrodt, F. D., Gervais, W. M., & Hilgard, J. (2019). Correcting for bias in psychology: A comparison of meta-analytic methods. *Advances in Methods and Practices in Psychological Science*, 2, 115–144. <http://dx.doi.org/10.1177/2515245919847196>
- Chan, J. C., Manley, K. D., Davis, S. D., & Szpunar, K. K. (2018). Testing potentiates new learning across a retention interval and a lag: A strategy change perspective. *Journal of Memory and Language*, 102, 83–96. <http://dx.doi.org/10.1016/j.jml.2018.05.007>
- Chan, J. C., Meissner, C. A., & Davis, S. D. (2018). Retrieval potentiates new learning: A theoretical and meta-analytic review. *Psychological Bulletin*, 144, 1111–1146. <http://dx.doi.org/10.1037/bul0000166>

- *Chan, P., Kim, S., Garavalia, L., & Wang, J. (2018). Implementing a strategy for promoting long-term meaningful learning in a pharmacokinetics course. *Currents in Pharmacy Teaching & Learning*, 10, 1048–1054. <http://dx.doi.org/10.1016/j.cptl.2018.05.013>
- *Chang, C. Y., Yeh, T. K., & Barufaldi, J. P. (2010). The positive and negative effects of science concept tests on student conceptual understanding. *International Journal of Science Education*, 32, 265–282. <http://dx.doi.org/10.1080/09500690802650055>
- *Chen, H.-Y., & Chuang, C.-H. (2012). The learning effectiveness of nursing students using online testing as an assistant tool: A cluster randomized controlled trial. *Nurse Education Today*, 32, 208–213. <http://dx.doi.org/10.1016/j.nedt.2011.03.004>
- *Cheng, C. K. (2014). *Effect of multiple-choice testing on memory retention–cue-target symmetry* (Doctoral dissertation). University of Toronto, Toronto, Canada. Retrieved from <https://tspace.library.utoronto.ca/handle/1807/65649>
- Cho, K. W., Neely, J. H., Crocco, S., & Vitrano, D. (2017). Testing enhances both encoding and retrieval for both tested and untested items. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, 70, 1–60. <http://dx.doi.org/10.1080/17470218.2016.1175485>
- *Chua, C., Foster, M., McKessock, D., & Smith, D. (2005). *The impact of frequent testing, online homework and contact hours on undergraduate achievement in business statistics*. Proceedings of the Annual Conference of the Administrative Sciences Association of Canada, Management Education Division.
- *Çil, E. (2015). Effect of two-tier diagnostic tests on promoting learners' conceptual understanding of variables in conducting scientific experiments. *Applied Measurement in Education*, 28, 253–273. <http://dx.doi.org/10.1080/08957347.2015.1064124>
- Coburn, K. M., & Vevea, J. L. (2019). *Estimating weight-function models for publication bias in R (R package version 2.0.1)*. Retrieved from <https://cran.r-project.org/web/packages/weightr/>
- *Cogliano, M., Kardash, C. M., & Bernacki, M. L. (2019). The effects of retrieval practice and prior topic knowledge on test performance and confidence judgments. *Contemporary Educational Psychology*, 56, 117–129. <http://dx.doi.org/10.1016/j.cedpsych.2018.12.001>
- *Cohonner, A., & Mayer, J. (2018). Retrieval-based learning in the context of inquiry-based learning. In N. Gericke & M. Grace (Eds.), *Challenges in biology education research* (pp. 273–287). Karlstad, Sweden: University Printing Office.
- *Coker, A. O., Lusk, K. A., Maize, D. F., Ramsinghani, S., Tabor, R. A., Yablonski, E. A., & Zertuche, A. (2018). The effect of repeated testing of pharmacy calculations and drug knowledge to improve knowledge retention in pharmacy students. *Currents in Pharmacy Teaching & Learning*, 10, 1609–1615. <http://dx.doi.org/10.1016/j.cptl.2018.08.019>
- *Coulter-Kern, R. G., Fogle, K. L., & Sibert, H. M. (2010). The effect of online quizzing on understanding of key concepts in an introduction to psychology course. *Journal of the Indiana Academy of the Social Sciences*, 14, 97–102.
- *Crossgrove, K., & Curran, K. L. (2008). Using clickers in nonmajors- and majors-level Biology courses: Student opinion, learning, and long-term retention of course material. *CBE Life Sciences Education*, 7, 146–154. <http://dx.doi.org/10.1187/cbe.07-08-0060>
- Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York, NY: Routledge.
- *Daigle, K. (2015). *Frequent quizzing versus class reviews* (Master's thesis). Louisiana State University, Baton Rouge, LA. Retrieved from https://digitalcommons.lsu.edu/gradschool_theses/692
- *Daniel, D. B., & Broida, J. (2004). Using web-based quizzing to improve exam performance: Lessons learned. *Teaching of Psychology*, 31, 207–208. http://dx.doi.org/10.1207/s15328023top3103_6
- *Davis, K. A. (2013). Using low- and no-stakes quizzing for student self-evaluation of readiness for exams. *International Journal of Construction Education and Research*, 9, 256–271. <http://dx.doi.org/10.1080/15578771.2013.809036>
- Davis, S. D., Chan, J. C. K., & Wilford, M. M. (2017). The dark side of interpolated testing: Frequent switching between retrieval and encoding impairs new learning. *Journal of Applied Research in Memory & Cognition*, 6, 434–441. <http://dx.doi.org/10.1016/j.jarmac.2017.07.002>
- *Delaram, M., Shams, S., & Gandomani, H. S. (2017). The effect of quizzes on test scores of nursing students for learning maternal and child health. *Journal of Medical Education*, 16, 118–122.
- *Denny, T., Paterson, J., & Feldhusen, J. (1964). Anxiety and achievement as functions of daily testing. *Journal of Educational Measurement*, 1, 143–147. <http://dx.doi.org/10.1111/j.1745-3984.1964.tb00172.x>
- *Dineen, P., Taylor, J., & Stephens, L. (1989). The effect of testing frequency upon the achievement of students in high school mathematics courses. *School Science and Mathematics*, 89, 197–200. <http://dx.doi.org/10.1111/j.1949-8594.1989.tb11910.x>
- *Dirkx, K. J. H., Kester, L., & Kirschner, P. A. (2014). The testing effect for learning principles and procedures from texts. *The Journal of Educational Research*, 107, 357–364. <http://dx.doi.org/10.1080/00220671.2013.823370>
- *Dobson, J. L. (2008). The use of formative online quizzes to enhance class preparation and scores on summative exams. *Advances in Physiology Education*, 32, 297–302. <http://dx.doi.org/10.1152/advan.90162.2008>
- *Dobson, J. L. (2013). Retrieval practice is an efficient method of enhancing the retention of anatomy and physiology information. *Advances in Physiology Education*, 37, 184–191. <http://dx.doi.org/10.1152/advan.00174.2012>
- *Dobson, J. L., & Linderholm, T. (2015). Self-testing promotes superior retention of anatomy and physiology information. *Advances in Health Sciences Education*, 20, 149–161. <http://dx.doi.org/10.1007/s10459-014-9514-8>
- *Dobson, J. L., Linderholm, T., & Perez, J. (2018). Retrieval practice enhances the ability to evaluate complex physiology information. *Medical Education*, 52, 513–525. <http://dx.doi.org/10.1111/medu.13503>
- *Dobson, J. L., Linderholm, T., & Stroud, L. (2019). Retrieval practice and judgements of learning enhance transfer of physiology information. *Advances in Health Sciences Education*, 24, 525–537. <http://dx.doi.org/10.1007/s10459-019-09881-w>
- *Dobson, J. L., Linderholm, T., & Yarbrough, M. B. (2015). Self-testing produces superior recall of both familiar and unfamiliar muscle information. *Advances in Physiology Education*, 39, 309–314. <http://dx.doi.org/10.1152/advan.00052.2015>
- *Dobson, J. L., Perez, J., & Linderholm, T. (2017). Distributed retrieval practice promotes superior recall of anatomy information. *Anatomical Sciences Education*, 10, 339–347. <http://dx.doi.org/10.1002/ase.1668>
- *Downs, S. D. (2015). Testing in the college classroom: Do testing and feedback influence grades throughout an entire semester? *Scholarship of Teaching and Learning in Psychology*, 1, 172–181. <http://dx.doi.org/10.1037/sti0000025>
- *Duchastel, P. C. (1981). Retention of prose following testing with different types of tests. *Contemporary Educational Psychology*, 6, 217–226. [http://dx.doi.org/10.1016/0361-476X\(81\)90002-3](http://dx.doi.org/10.1016/0361-476X(81)90002-3)
- *Duchastel, P. C., & Nungester, R. J. (1982). Testing effects measured with alternate test forms. *The Journal of Educational Research*, 75, 309–313. <http://dx.doi.org/10.1080/00220671.1982.10885400>
- Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, 14, 4–58. <http://dx.doi.org/10.1177/1529100612453266>
- Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, 56, 455–463. <http://dx.doi.org/10.1111/j.0006-341X.2000.00455.x>

- Dynarski, S. M. (2017). *For better learning in college lectures, lay down the laptop and pick up a pen*. Retrieved from <https://www.brookings.edu/research/for-better-learning-in-college-lectures-lay-down-the-laptop-and-pick-up-a-pen/>
- Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal*, 315, 629–634. <http://dx.doi.org/10.1136/bmj.315.7109.629>
- *Einstein, G. O., Mullet, H. G., & Harrison, T. L. (2012). The testing effect: Illustrating a fundamental concept and changing study strategies. *Teaching of Psychology*, 39, 190–193. <http://dx.doi.org/10.1177/0098628312450432>
- Eriksson, J., Kalpouzos, G., & Nyberg, L. (2011). Rewiring the brain with repeated retrieval: A parametric fMRI study of the testing effect. *Neuroscience Letters*, 505, 36–40. <http://dx.doi.org/10.1016/j.neulet.2011.08.061>
- *Eustace, E., Bradford, M., & Pathak, P. (2015, October). *A practice testing learning framework to enhance transfer in mathematics*. Paper presented at the 14th IT&T Conference 2015, National College of Ireland, Ireland. http://ittconference.ie/uploads/conference/2015/ITT_Conference_Proceedings_2015.pdf#page=95
- *Evans, D. D. (2013). *Quizzing and retention in the high school science class* (Master's thesis). Louisiana State University, Baton Rouge, LA. Retrieved from https://digitalcommons.lsu.edu/gradschool_theses/3780/
- Fajnzylber, E., Lara, B., & León, T. (2019). Increased learning or GPA inflation? Evidence from GPA-based university admission in Chile. *Economics of Education Review*, 72, 147–165. <http://dx.doi.org/10.1016/j.econedurev.2019.05.009>
- Fazio, L. K., Agarwal, P. K., Marsh, E. J., & Roediger, H. L. (2010). Memorial consequences of multiple-choice testing on immediate and delayed tests. *Memory & Cognition*, 38, 407–418. <http://dx.doi.org/10.3758/MC.38.4.407>
- *Feldman, M., Fernando, O., Wan, M., Martimianakis, M. A., & Kulasegaram, K. (2018). Testing test-enhanced continuing medical education: A randomized controlled trial. *Academic Medicine*, 93, S30–S36. <http://dx.doi.org/10.1097/ACM.0000000000002377>
- Fey, L. (2012). *'Drill and kill' testing: Just say "no"*. Retrieved from <https://www.msdf.org/blog/2012/03/drill-and-kill-testing-just-say-no/>
- Fisher, Z., & Tipton, E. (2015). *robumeta: An R-package for robust variance estimation in meta-analysis*. Retrieved from <https://arxiv.org/abs/1503.02220>
- *Foss, D. J., & Pirozzolo, J. W. (2017). Four semesters investigating frequency of testing, the testing effect, and transfer of training. *Journal of Educational Psychology*, 109, 1067–1083. <http://dx.doi.org/10.1037/edu000197>
- *Gauci, S. A., Dantas, A. M., Williams, D. A., & Kemm, R. E. (2009). Promoting student-centered active learning in lectures with a personal response system. *Advances in Physiology Education*, 33, 60–71. <http://dx.doi.org/10.1152/advan.00109.2007>
- *Gay, L. R., & Gallagher, P. D. (1976). The comparative effectiveness of tests versus written exercises 1. *The Journal of Educational Research*, 70, 59–61. <http://dx.doi.org/10.1080/00220671.1976.10884951>
- *Gaynor, J., & Millham, J. (1976). Student performance and evaluation under variant teaching and testing methods in a large college course. *Journal of Educational Psychology*, 68, 312–317. <http://dx.doi.org/10.1037/0022-0663.68.3.312>
- *Geiger, O. G., & Bostow, D. E. (1976). Contingency-managed college instruction: Effects of weekly quizzes on performance on examination. *Psychological Reports*, 39, 707–710. <http://dx.doi.org/10.2466/pr0.1976.39.3.707>
- *Geller, J., Carpenter, S. K., Lamm, M. H., Rahman, S., Armstrong, P. I., & Coffman, C. R. (2017). Prequestions do not enhance the benefits of retrieval in a STEM classroom. *Cognitive Research: Principles and Implications*, 2, 42. <http://dx.doi.org/10.1186/s41235-017-0078-z>
- Geller, J., Toftness, A. R., Armstrong, P. I., Carpenter, S. K., Manz, C. L., Coffman, C. R., & Lamm, M. H. (2018). Study strategies and beliefs about learning as a function of academic achievement and achievement goals. *Memory*, 26, 683–690. <http://dx.doi.org/10.1080/09658211.2017.1397175>
- Gervais, W. (2015). *Putting PET-PEESE to the test*. Retrieved from <http://crystalprisonzone.blogspot.com/2015/06/putting-pet-peese-to-test-part-1a.html>
- *Gholami, V., & Moghaddam, M. M. (2013). The effect of weekly quizzes on students' final achievement score. *International Journal of Modern Education and Computer Science*, 5, 36–41. <http://dx.doi.org/10.5815/ijmecs.2013.01.05>
- *Ghorbani, M. R. (2017). Quizzes in every other session improve undergraduate EFL learners' pronunciation achievement. *Advances in Language and Literary Studies*, 8, 65–70. <http://dx.doi.org/10.7575/aiac.all.v.8n.5p.65>
- *Gier, V. S., & Kreiner, D. S. (2009). Incorporating active learning with PowerPoint-based lectures using content-based questions. *Teaching of Psychology*, 36, 134–139. <http://dx.doi.org/10.1080/00986280902739792>
- *Glass, A. L., Brill, G., & Ingate, M. (2008). Combined online and in-class pretesting improves exam performance in general psychology. *Educational Psychology*, 28, 483–503. <http://dx.doi.org/10.1080/01443410701777280>
- *Glavin, D. (2012). *The relationship of quizzing and student success in a college level core statistics course* (Master's thesis). University of New Mexico, Albuquerque, NM. Retrieved from https://digitalrepository.unm.edu/math_etds/78
- Glover, J. A. (1989). The "testing" phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology*, 81, 392–399. <http://dx.doi.org/10.1037/0022-0663.81.3.392>
- *Gokcora, D., & DePaulo, D. (2018). Frequent quizzes and student improvement of reading: A pilot study in a community college setting. *SAGE Open*, 8, 2. <http://dx.doi.org/10.1177/2158244018782580>
- 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. <http://dx.doi.org/10.1177/0098628312450436>
- *Goossens, N. A. M. C., Camp, G., Verkoeijen, P. P. J. L., Tabbers, H. K., Bouwmeester, S., & Zwaan, R. A. (2016). Distributed practice and retrieval practice in primary school vocabulary learning: A multi-classroom study. *Applied Cognitive Psychology*, 30, 700–712. <http://dx.doi.org/10.1002/acp.3245>
- *Goossens, N. A. M. C., Camp, G., Verkoeijen, P. P. J. L., Tabbers, H. K., & Zwaan, R. A. (2014). The benefit of retrieval practice over elaborative restudy in primary school vocabulary learning. *Journal of Applied Research in Memory & Cognition*, 3, 177–182. <http://dx.doi.org/10.1016/j.jarmac.2014.05.003>
- *Graham, R. B. (1999). Unannounced quizzes raise test scores selectively for mid-range students. *Teaching of Psychology*, 26, 271–273. <http://dx.doi.org/10.1207/S15328023TOP260406>
- Gravetter, F. J., & Forzano, L. A. (2011). *Research methods for the behavioral sciences* (4th ed.). Boston, MA: Cengage Learning.
- *Grimstad, K., & Grabe, M. (2004). Are online study questions beneficial? *Teaching of Psychology*, 31, 143–146. http://dx.doi.org/10.1207/s15328023top3102_8
- *Gunasekera, T. W. (1997). *Effects of pretest sensitization associated with cooperative learning strategies on the achievement level of adult mathematics students* (Doctoral dissertation). Wayne State University, Detroit, MI. Retrieved from <http://libproxy1.nus.edu.sg/login?url=https://search-proquest-com.libproxy1.nus.edu.sg/docview/304378570?accountid=13876>
- Haberyan, A., & Barnett, J. (2010). Collaborative testing and achievement: Are two heads really better than one? *Journal of Instructional Psychology*, 37, 32–41.

- *Haberyan, K. A. (2003). Do weekly quizzes improve student performance on general biology exams? *The American Biology Teacher*, 65, 110–114. <http://dx.doi.org/10.2307/4451449>
- *Harrington, M., & Jiang, W. (2013). Focus on the forms: Form recognition practice in Chinese vocabulary learning. *Australian Review of Applied Linguistics*, 36, 132–145. <http://dx.doi.org/10.1075/aral.36.2.01har>
- Hartwig, M. K., & Dunlosky, J. (2012). Study strategies of college students: Are self-testing and scheduling related to achievement? *Psychonomic Bulletin & Review*, 19, 126–134. <http://dx.doi.org/10.3758/s13423-011-0181-y>
- Hattie, J. A. C. (2009). *Visible learning: A synthesis of 800+ meta-analyses on achievement*. London, UK: Routledge.
- *Haynie, W. J. (1994). Effect of multiple choice and short answer tests on delayed retention learning. *Journal of Technology Education*, 6, 32–44. <http://dx.doi.org/10.21061/jte.v6i1.a.3>
- *Haynie, W. J. (2003). Effects of multiple-choice and matching tests on delayed retention learning in postsecondary metals technology. *Journal of STEM Teacher Education*, 4, 7–22.
- Healy, A. F., Jones, M., Lalchandani, L., & Tack, L. A. (2017). Timing of quizzes during learning: Effects on motivation and retention. *Journal of Experimental Psychology: Applied*, 23, 128–137. <http://dx.doi.org/10.1037/xap0000123>
- Hedges, L. V. (1982). Estimation of effect size from a series of independent experiments. *Psychological Bulletin*, 92, 490–499. <http://dx.doi.org/10.1037/0033-2909.92.2.490>
- Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, 1, 39–65. <http://dx.doi.org/10.1002/jrsm.5>
- Heiner, C. E., Banet, A. I., & Wieman, C. (2014). Preparing students for class: How to get 80% of students reading the textbook before class. *American Journal of Physics*, 82, 989–996. <http://dx.doi.org/10.1119/1.4895008>
- *Heinicke, M. R., Zuckerman, C. K., & Cravalho, D. A. (2017). An evaluation of readiness assessment tests in a college classroom: Exam performance, attendance, and participation. *Behavior Analysis: Research and Practice*, 17, 129–141. <http://dx.doi.org/10.1037/bar0000073>
- Heitmann, S., Grund, A., Berthold, K., Fries, S., & Roelle, J. (2018). Testing is more desirable when it is adaptive and still desirable when compared to note-taking. *Frontiers in Psychology*, 9, 2596. <http://dx.doi.org/10.3389/fpsyg.2018.02596>
- Hembree, R. (1988). Correlates, causes, effects, and treatment of test anxiety. *Review of Educational Research*, 58, 47–77. <http://dx.doi.org/10.3102/00346543058001047>
- *Hennig, S., Staatz, C. E., Bond, J. A., Leung, D., & Singleton, J. (2019). Quizzing for success: Evaluation of the impact of feedback quizzes on the experiences and academic performance of undergraduate students in two clinical pharmacokinetics courses. *Currents in Pharmacy Teaching & Learning*, 11, 742–749. <http://dx.doi.org/10.1016/j.cptl.2019.03.014>
- Herlitz, A., Nilsson, L.-G., & Bäckman, L. (1997). Gender differences in episodic memory. *Memory & Cognition*, 25, 801–811. <http://dx.doi.org/10.3758/BF03211324>
- *Hesse, R. M. (1971). *The effect of daily quizzes on hour examination performance in a junior level psychology course* (Master's thesis). Western Michigan University, Kalamazoo, MI. Retrieved from https://scholarworks.wmich.edu/masters_theses
- Hilgard, J. (2017). *Trim-and-fill just doesn't work*. Retrieved from <http://crystalprisonzone.blogspot.com/2017/05/trim-and-fill-just-doesnt-work.html>
- Hilgard, J., Sala, G., Boot, W. R., & Simons, D. J. (2019). Overestimation of action-game training effects: Publication bias and salami slicing. *Collabra Psychology*, 5, 30. <http://dx.doi.org/10.1525/collabra.231>
- *Hirschman, B. (2017). *The effects of daily quizzes on student achievement in a chemistry class* (Master's thesis). Montana State University, Bozeman, MT. Retrieved from <https://scholarworks.montana.edu/xmlui/bitstream/handle/1/13665/HirschmanB0817.pdf?sequence=3>
- *Howard, C. R. (2010). *Examining the testing effect in an introductory psychology course* (Doctoral dissertation). Auburn University, Auburn, AL. Retrieved from <http://hdl.handle.net/10415/2289>
- *Howe, P. D. L., McKague, M., Lodge, J. M., Blunden, A. G., & Saw, G. (2018). PeerWise: Evaluating the effectiveness of a web-based learning aid in a second-year psychology subject. *Psychology Learning & Teaching*, 17, 166–176. <http://dx.doi.org/10.1177/1475725718764181>
- *Inouye, C. Y., Bae, C. L., & Hayes, K. N. (2017). Using whiteboards to support college students' learning of complex physiological concepts. *Advances in Physiology Education*, 41, 478–484. <http://dx.doi.org/10.1152/advan.00202.2016>
- *Iwamoto, D. H., Hargis, J., Taitano, E. J., & Vuong, K. (2017). Analyzing the efficacy of the testing effect using Kahoot™ on student performance. *Turkish Online Journal of Distance Education*, 18, 80–93. <http://dx.doi.org/10.17718/tojde.306561>
- Jacoby, L. L., Wahlheim, C. N., & Coane, J. H. (2010). Test-enhanced learning of natural concepts: Effects on recognition memory, classification, and metacognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 36, 1441–1451. <http://dx.doi.org/10.1037/a0020636>
- *Jaeger, A., Eisenkraemer, R. E., & Stein, L. M. (2015). Test-enhanced learning in third-grade children. *Educational Psychology*, 35, 513–521. <http://dx.doi.org/10.1080/01443410.2014.963030>
- *Jägerskog, A.-S. (2015). *Pictures and a thousand words: Learning psychology through visual illustrations and testing* (Doctoral dissertation). Stockholm University, Stockholm, Sweden. Retrieved from <http://www.diva-portal.org/smash/record.jsf?pid=diva2%3A883530&dsid=4598>
- *Jägerskog, A.-S., Jönsson, F. U., Selander, S., & Jonsson, B. (2019). Multimedia learning trumps retrieval practice in psychology teaching. *Scandinavian Journal of Psychology*, 60, 222–230. <http://dx.doi.org/10.1111/sjop.12527>
- *Janczarek, K. (1970). *The effect of daily quizzes on hour exam performance* (Masters thesis). Western Michigan University, Kalamazoo, MI. Retrieved from https://scholarworks.wmich.edu/masters_theses
- Jing, H. G., Szpunar, K. K., & Schacter, D. L. (2016). Interpolated testing influences focused attention and improves integration of information during a video-recorded lecture. *Journal of Experimental Psychology: Applied*, 22, 305–318. <http://dx.doi.org/10.1037/xap0000087>
- *Johnson, B. E. (1938). The effect of written examinations on learning and on the retention of learning. *Journal of Experimental Education*, 7, 55–62. <http://dx.doi.org/10.1080/00220973.1938.11010116>
- Johnson, B. C., & Kiviniemi, M. T. (2009). The effect of online chapter quizzes on exam performance in an undergraduate social psychology course. *Teaching of Psychology*, 36, 33–37. <http://dx.doi.org/10.1080/00986280802528972>
- *Johnson, D. I., & Mrowka, K. (2010). Generative learning, quizzing and cognitive learning: An experimental study in the communication classroom. *Communication Education*, 59, 107–123. <http://dx.doi.org/10.1080/03634520903524739>
- *Johnson, M. C. (2012). *Pretesting in science: Effect on unit test scores* (Master's thesis). Louisiana State University, Baton Rouge, LA. Retrieved from https://digitalcommons.lsu.edu/gradschool_theses/3273
- *Johnson, P. E. (1990). Effect of frequent testing on learning mathematics. *International Journal of Mathematical Education in Science and Technology*, 21, 733–737. <http://dx.doi.org/10.1080/00207399000210507>
- *Jones, A. C., Wardlow, L., Pan, S. C., Zepeda, C., Heyman, G. D., Dunlosky, J., & Rickard, T. C. (2016). Beyond the rainbow: Retrieval practice leads to better spelling than does rainbow writing. *Educational Psychology Review*, 28, 385–400. <http://dx.doi.org/10.1007/s10648-015-9330-6>
- Kang, S. H. K., McDermott, K. B., & Roediger, H. L., III. (2007). Test format and corrective feedback modify the effect of testing on long-term

- retention. *European Journal of Cognitive Psychology*, 19, 528–558. <http://dx.doi.org/10.1080/09541440601056620>
- Kang, S. H. K., & Pashler, H. (2014). Is the benefit of retrieval practice modulated by motivation? *Journal of Applied Research in Memory & Cognition*, 3, 183–188. <http://dx.doi.org/10.1016/j.jarmac.2014.05.006>
- Karpicke, J. D. (2017). Retrieval-based learning: A decade of progress. In J. H. Byrne (Ed.), *Cognitive psychology of memory*, Vol. 2 of *Learning and memory: A comprehensive reference* (pp. 1–26). Amsterdam, the Netherlands: Elsevier. <http://dx.doi.org/10.1016/B978-0-12-809324-5.21055-9>
- Karpicke, J. D., & Bauernschmidt, A. (2011). Spaced retrieval: Absolute spacing enhances learning regardless of relative spacing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37, 1250–1257. <http://dx.doi.org/10.1037/a0023436>
- Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science*, 331, 772–775. <http://dx.doi.org/10.1126/science.1199327>
- *Karpicke, J. D., Blunt, J. R., & Smith, M. A. (2016). Retrieval-based learning: Positive effects of retrieval practice in elementary school children. *Frontiers in Psychology*. Advance online publication. <http://dx.doi.org/10.3389/fpsyg.2016.00350>
- *Karpicke, J. D., Blunt, J. R., Smith, M. A., & Karpicke, S. S. (2014). Retrieval-based learning: The need for guided retrieval in elementary school children. *Journal of Applied Research in Memory & Cognition*, 3, 198–206. <http://dx.doi.org/10.1016/j.jarmac.2014.07.008>
- 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. <http://dx.doi.org/10.1080/09658210802647009>
- Karpicke, J. D., Lehman, M., & Aue, W. R. (2014). Retrieval-based learning: An episodic context account. *Psychology of Learning and Motivation*, 61, 237–284. <http://dx.doi.org/10.1016/B978-0-12-800283-4.00007-1>
- Karpicke, J. D., & Roediger, H. L. (2007a). Expanding retrieval practice promotes short-term retention, but equally spaced retrieval enhances long-term retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33, 704–719. <http://dx.doi.org/10.1037/0278-7393.33.4.704>
- Karpicke, J. D., & Roediger, H. L., III. (2007b). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language*, 57, 151–162. <http://dx.doi.org/10.1016/j.jml.2006.09.004>
- Karpicke, J. D., & Roediger, H. L., III. (2008). The critical importance of retrieval for learning. *Science*, 319, 966–968. <http://dx.doi.org/10.1126/science.1152408>
- *Kelley, M. R., Chapman-Orr, E. K., Calkins, S., & Lemke, R. J. (2019). Generation and retrieval practice effects in the classroom using Peer-Wise. *Teaching of Psychology*, 46, 121–126. <http://dx.doi.org/10.1177/0098628319834174>
- *Keys, N. (1934). The influence on learning and retention of weekly as opposed to monthly tests. *Journal of Educational Psychology*, 25, 427–436. <http://dx.doi.org/10.1037/h0074468>
- *Khanna, M. M. (2015). Ungraded pop quizzes: Test-enhanced learning without all the anxiety. *Teaching of Psychology*, 42, 174–178. <http://dx.doi.org/10.1177/0098628315573144>
- *Khanna, M. M., & Cortese, M. J. (2016). The benefits of quizzing in content-focused versus skills-focused courses. *Scholarship of Teaching and Learning in Psychology*, 2, 87–97. <http://dx.doi.org/10.1037/stl0000051>
- *Kilickaya, F. (2017). The effects of pre-lecture online quizzes on language students' perceived preparation and academic performance. *PASAA: Journal of Language Teaching and Learning*, 53, 59–84.
- *King, A. (1992). Comparison of self-questioning, summarizing, and notetaking-review as strategies for learning from lectures. *American Educational Research Journal*, 29, 303–323. <http://dx.doi.org/10.3102/00028312029002303>
- Kirk-Johnson, A., Galla, B. M., & Fraundorf, S. H. (2019). Perceiving effort as poor learning: The misinterpreted-effort hypothesis of how experienced effort and perceived learning relate to study strategy choice. *Cognitive Psychology*, 115, 101237. <http://dx.doi.org/10.1016/j.cogpsych.2019.101237>
- *Klass, G., & Crothers, L. (2000). An experimental evaluation of web-based tutorial quizzes. *Social Science Computer Review*, 18, 508–515. <http://dx.doi.org/10.1177/089443930001800413>
- Kling, N., McCorkle, D., Miller, C., & Reardon, J. (2005). The impact of testing frequency on student performance in a marketing course. *Journal of Education for Business*, 81, 67–72. <http://dx.doi.org/10.3200/JOEB.81.2.67-72>
- Kornell, N., & Bjork, R. A. (2007). The promise and perils of self-regulated study. *Psychonomic Bulletin & Review*, 14, 219–224. <http://dx.doi.org/10.3758/BF03194055>
- Kornell, N., & Bjork, R. A. (2008). Optimising self-regulated study: The benefits—and costs—of dropping flashcards. *Memory*, 16, 125–136. <http://dx.doi.org/10.1080/09658210701763899>
- *Kromann, C. B., Bohnstedt, C., Jensen, M. L., & Ringsted, C. (2010). The testing effect on skills learning might last 6 months. *Advances in Health Sciences Education*, 15, 395–401. <http://dx.doi.org/10.1007/s10459-009-9207-x>
- *Kromann, C. B., Jensen, M. L., & Ringsted, C. (2009). The effect of testing on skills learning. *Medical Education*, 43, 21–27. <http://dx.doi.org/10.1111/j.1365-2923.2008.03245.x>
- *Kromann, C. B., Jensen, M. L., & Ringsted, C. (2011). Test-enhanced learning may be a gender-related phenomenon explained by changes in cortisol level. *Medical Education*, 45, 192–199. <http://dx.doi.org/10.1111/j.1365-2923.2010.03790.x>
- *Kulkarni, P. P., & Kulkarni, P. P. (2018). Post lecture multiple choice questions (MCQs) test improves the performance of students. *International Journal of Current Research and Review*, 10, 1–5. <http://dx.doi.org/10.31782/IJCRR.2018.10181>
- *LaDisa, A. G., & Biesboer, A. (2017). Incorporation of practice testing to improve knowledge acquisition in a pharmacotherapy course. *Currents in Pharmacy Teaching & Learning*, 9, 660–665. <http://dx.doi.org/10.1016/j.cptl.2017.03.002>
- *Lambert, T. (2010). *Interteaching and the testing effect: How quizzes alter the efficacy of interteaching* (Master's thesis). James Madison University, Harrisonburg, VA. Retrieved from <https://commons.lib.jmu.edu/master201019/397/>
- Lantz, M. E., & Stawiski, A. (2014). Effectiveness of clickers: Effect of feedback and the timing of questions on learning. *Computers in Human Behavior*, 31, 280–286. <http://dx.doi.org/10.1016/j.chb.2013.10.009>
- *Larsen, D. P., Butler, A. C., Aung, W. Y., Corboy, J. R., Friedman, D. I., & Sperling, M. R. (2015). The effects of test-enhanced learning on long-term retention in AAN annual meeting courses. *Neurology*, 84, 748–754. <http://dx.doi.org/10.1212/WNL.0000000000001264>
- *Larsen, D. P., Butler, A. C., Lawson, A. L., & Roediger, H. L., III. (2013). The importance of seeing the patient: Test-enhanced learning with standardized patients and written tests improves clinical application of knowledge. *Advances in Health Sciences Education*, 18, 409–425. <http://dx.doi.org/10.1007/s10459-012-9379-7>
- *Larsen, D. P., Butler, A. C., & Roediger, H. L., III. (2013). Comparative effects of test-enhanced learning and self-explanation on long-term retention. *Medical Education*, 47, 674–682. <http://dx.doi.org/10.1111/medu.12141>
- *Larson, E. D. (2017). *The effects of iClicker bar graph feedback on test performance* (Master's thesis). Montana State University, Bozeman, MT. Retrieved from <https://scholarworks.montana.edu/xmlui/handle/1/12792>

- Lazowski, R. A., & Hulleman, C. S. (2016). Motivation interventions in education: A meta-analytic review. *Review of Educational Research*, 86, 602–640. <http://dx.doi.org/10.3102/0034654315617832>
- *Leahy, W., Hanham, J., & Sweller, J. (2015). High element interactivity information during problem solving may lead to failure to obtain the testing effect. *Educational Psychology Review*, 27, 291–304. <http://dx.doi.org/10.1007/s10648-015-9296-4>
- Lechuga, M. T., Ortega-Tudela, J. M., & Gómez-Ariza, C. J. (2015). Further evidence that concept mapping is not better than repeated retrieval as a tool for learning from texts. *Learning and Instruction*, 40, 61–68. <http://dx.doi.org/10.1016/j.learninstruc.2015.08.002>
- Lee, H. S., & Ahn, D. (2018). Testing prepares students to learn better: The forward effect of testing in category learning. *Journal of Educational Psychology*, 110, 203–217. <http://dx.doi.org/10.1037/edu0000211>
- *Leeming, F. C. (2002a). The Exam-A-Day procedure improves performance in psychology classes. *Teaching of Psychology*, 29, 210–212. http://dx.doi.org/10.1207/S15328023TOP2903_06
- Leeming, F. C. (2002b). Retrieving essential material at the end of lectures improves performance on statistics exams. *Teaching of Psychology*, 38, 94–97. <http://dx.doi.org/10.1177/0098628311401587>
- *Leggett, J. M. I., Burt, J. S., & Carroll, A. (2019). Retrieval practice can improve classroom review despite low practice test performance. *Applied Cognitive Psychology*, 33, 759–770. <http://dx.doi.org/10.1002/acp.3517>
- Leight, H., Saunders, C., Calkins, R., & Withers, M. (2012). Collaborative testing improves performance but not content retention in a large-enrollment introductory biology class. *CBE Life Sciences Education*, 11, 392–401. <http://dx.doi.org/10.1187/cbe.12-04-0048>
- Lewin, C., & Herlitz, A. (2002). Sex differences in face recognition—Women's faces make the difference. *Brain and Cognition*, 50, 121–128. [http://dx.doi.org/10.1016/S0278-2626\(02\)00016-7](http://dx.doi.org/10.1016/S0278-2626(02)00016-7)
- *Lin, T. (2016). Effects of different types of quizzes on student effort investment behaviour and learning outcomes. *International Journal of Education Economics and Development*, 7, 33–52. <http://dx.doi.org/10.1504/IJEED.2016.079238>
- *Linderholm, T., Dobson, J., & Yarbrough, M. B. (2016). The benefit of self-testing and interleaving for synthesizing concepts across multiple physiology texts. *Advances in Physiology Education*, 40, 329–334. <http://dx.doi.org/10.1152/advan.00157.2015>
- *Lipko-Speed, A., Dunlosky, J., & Rawson, K. A. (2014). Does testing with feedback help grade-school children learn key concepts in science? *Journal of Applied Research in Memory & Cognition*, 3, 171–176. <http://dx.doi.org/10.1016/j.jarmac.2014.04.002>
- Little, J. L., Storm, B. C., & Bjork, E. L. (2011). The costs and benefits of testing text materials. *Memory*, 19, 346–359. <http://dx.doi.org/10.1080/09658211.2011.569725>
- *Lloyd, E. P., Walker, R. J., Metz, M. A., & Diekman, A. B. (2018). Comparing review strategies in the classroom: Self-testing yields more favorable student outcomes relative to question generation. *Teaching of Psychology*, 45, 115–123. <http://dx.doi.org/10.1177/0098628318762871>
- Lortie-Forgues, H., & Inglis, M. (2019). Rigorous large-scale educational RCTs are often uninformative: Should we be concerned? *Educational Researcher*, 48, 158–166. <http://dx.doi.org/10.3102/0013189X19832850>
- Lundeberg, M. A., & Fox, P. W. (1991). Do laboratory findings on test expectancy generalize to classroom outcomes? *Review of Educational Research*, 61, 94–106. <http://dx.doi.org/10.3102/00346543061001094>
- *Lyle, K. B., & Crawford, N. A. (2011). Retrieving essential material at the end of lectures improves performance on statistics exams. *Teaching of Psychology*, 38, 94–97. <http://dx.doi.org/10.1177/0098628311401587>
- *Ma, X. (1995). The effect of informal oral testing frequency upon mathematics learning of high school students in China. *Journal of Classroom Interaction*, 30, 17–20.
- *Mains, T. E., Cofrancesco, J., Jr., Milner, S. M., Shah, N. G., & Goldberg, H. (2015). Do questions help? The impact of audience response systems on medical student learning: A randomised controlled trial. *Postgraduate Medical Journal*, 91, 361–367. <http://dx.doi.org/10.1136/postgradmedj-2014-132987>
- *Makarchuk, D. (2018). Recall Efficacy in EFL Learning. *English Teaching*, 73, 115–138. <http://dx.doi.org/10.15858/engtea.73.2.201806.115>
- *Maloney, E. L., & Ruch, G. M. (1929). The use of objective tests in teaching as illustrated by grammar. *The School Review*, 37, 62–66. <http://dx.doi.org/10.1086/438790>
- *Marden, N. Y., Ulman, L. G., Wilson, F. S., & Velan, G. M. (2013). Online feedback assessments in physiology: Effects on students' learning experiences and outcomes. *Advances in Physiology Education*, 37, 192–200. <http://dx.doi.org/10.1152/advan.00092.2012>
- Martin, D., Friesen, E., & De Pau, A. (2014). Three heads are better than one: A mixed methods study examining collaborative versus traditional test-taking with nursing students. *Nurse Education Today*, 34, 971–977. <http://dx.doi.org/10.1016/j.nedt.2014.01.004>
- *Martin, R. R., & Srikameswaran, K. (1974). Correlation between frequent testing and student performance. *Journal of Chemical Education*, 51, 485–486. <http://dx.doi.org/10.1021/ed051p485>
- *Maurer, T. W. (2006). Daily online extra credit quizzes and exam performance. *Journal of Teaching in Marriage & Family*, 6, 227–238. Retrieved from <https://digitalcommons.georgiasouthern.edu/ecology-facpubs/18>
- *Mayer, R. E., Stull, A., DeLeeuw, K., Almeroth, K., Bimber, B., Chun, D., . . . Zhang, H. (2009). Clickers in college classrooms: Fostering learning with questioning methods in large lecture classes. *Contemporary Educational Psychology*, 34, 51–57. <http://dx.doi.org/10.1016/j.cedpsych.2008.04.002>
- McAndrew, M., Morrow, C. S., Atiyeh, L., & Pierre, G. C. (2016). Dental student study strategies: Are self-testing and scheduling related to academic performance? *Journal of Dental Education*, 80, 542–552. <http://dx.doi.org/10.1002/j.0022-0337.2016.80.5.tb06114.x>
- *McConnell, M., Azzam, K., Xenodemetropoulos, T., & Panju, A. (2015). Effectiveness of test-enhanced learning in continuing health sciences education: A randomized controlled trial. *The Journal of Continuing Education in the Health Professions*, 35, 119–122. <http://dx.doi.org/10.1002/chp.21293>
- *McConnell, M., Hou, C., Panju, M., Panju, A., & Azzam, K. (2018). Does testing enhance learning in continuing medical education? *Canadian Medical Education Journal*, 9, e83–e88. <http://dx.doi.org/10.36834/cmej.42236>
- *McConnell, M., St-Onge, C., & Young, M. E. (2015). The benefits of testing for learning on later performance. *Advances in Health Sciences Education*, 20, 305–320. <http://dx.doi.org/10.1007/s10459-014-9529-1>
- *McDaniel, M. A., Agarwal, P. K., Huelser, B. J., McDermott, K. B., & Roediger, H. L. (2011). Test-enhanced learning in a middle school science classroom: The effects of quiz frequency and placement. *Journal of Educational Psychology*, 103, 399–414. <http://dx.doi.org/10.1037/a0021782>
- *McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology*, 19, 494–513. <http://dx.doi.org/10.1080/09541440701326154>
- *McDaniel, M. A., McDermott, K. B., Agarwal, P. K., & Roediger, H. L. (2008, June). *Test-enhanced learning in the classroom: The Columbia Middle School project*. Paper presented at the Meeting of the Institute of Education Sciences Research, Washington, DC. Retrieved from <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.535.9415&rep=rep1&type=pdf>
- *McDaniel, M. A., Thomas, R. C., Agarwal, P. K., McDermott, K. B., & Roediger, H. L. (2013). Quizzing in middle-school science: Successful

- transfer performance on classroom exams. *Applied Cognitive Psychology*, 27, 360–372. <http://dx.doi.org/10.1002/acp.2914>
- McDermott, K. B. (2006). Paradoxical effects of testing: Repeated retrieval attempts enhance the likelihood of later accurate and false recall. *Memory & Cognition*, 34, 261–267. <http://dx.doi.org/10.3758/BF03193404>
- *McDermott, K. B., Agarwal, P. K., D'Antonio, L., Roediger, H. L., III, & McDaniel, M. A. (2014). Both multiple-choice and short-answer quizzes enhance later exam performance in middle and high school classes. *Journal of Experimental Psychology: Applied*, 20, 3–21. <http://dx.doi.org/10.1037/xap0000004>
- McShane, B. B., Böckenholt, U., & Hansen, K. T. (2016). Adjusting for publication bias in meta-analysis: An evaluation of selection methods and some cautionary notes. *Perspectives on Psychological Science*, 11, 730–749. <http://dx.doi.org/10.1177/1745691616662243>
- *Mendes, P. S. O. (2015). *Testing effect on a college course: Examining test format and repeated questioning*. Universidade do Minho. Retrieved from <https://repositorium.sdum.uminho.pt/handle/1822/37597>
- *Messineo, L., Gentile, M., & Allegra, M. (2015). Test-enhanced learning: Analysis of an experience with undergraduate nursing students. *BMC Medical Education*, 15, 182–189. <http://dx.doi.org/10.1186/s12909-015-0464-5>
- *Michaels, J. L. (2017). Quizzes benefit freshman and sophomore students more than junior and senior students in introductory psychology classes with noncumulative exams. *Scholarship of Teaching and Learning in Psychology*, 3, 272–283. <http://dx.doi.org/10.1037/stl0000098>
- *Mok, W. S. Y., & Chan, W. W. L. (2016). How do tests and summary writing tasks enhance long-term retention of students with different levels of test anxiety? *Instructional Science*, 44, 567–581. <http://dx.doi.org/10.1007/s11251-016-9393-x>
- Morehead, K., Rhodes, M. G., & DeLozier, S. (2016). Instructor and student knowledge of study strategies. *Memory*, 24, 257–271. <http://dx.doi.org/10.1080/09658211.2014.1001992>
- Moreira, B. F. T., Pinto, T. S. S., Starling, D. S. V., & Jaeger, A. (2019). Retrieval practice in classroom settings: A review of applied research. *Frontiers in Psychology*, 4, 5. <http://dx.doi.org/10.3389/fpsyg.2019.00005>
- *Morling, B., McAuliffe, M., Cohen, L., & DiLorenzo, T. M. (2008). Efficacy of personal response systems (“clickers”) in large, introductory psychology classes. *Teaching of Psychology*, 35, 45–50. <http://dx.doi.org/10.1177/009862830803500112>
- 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. [http://dx.doi.org/10.1016/S0022-5371\(77\)80016-9](http://dx.doi.org/10.1016/S0022-5371(77)80016-9)
- Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods*, 7, 105–125. <http://dx.doi.org/10.1037/1082-989X.7.1.105>
- Murphy, D. P., & Stanga, K. G. (1994). The effects of frequent testing in an income tax course: An experiment. *Journal of Accounting Education*, 12, 27–41. [http://dx.doi.org/10.1016/0748-5751\(94\)90017-5](http://dx.doi.org/10.1016/0748-5751(94)90017-5)
- *Nakos, G., & Whiting, A. (2018). The role of frequent short exams in improving student performance in hybrid global business classes. *Journal of Education for Business*, 93, 51–57. <http://dx.doi.org/10.1080/08832323.2017.1417231>
- *Narloch, R., Garbin, C. P., & Turnage, K. D. (2006). Benefits of prelecture quizzes. *Teaching of Psychology*, 33, 109–112. http://dx.doi.org/10.1207/s15328023top3302_6
- *Negin, G. A. (1981). The effects of test frequency in a first-year torts course. *Journal of Legal Education*, 31, 673–676. Retrieved from <http://www.jstor.org/stable/42892366>
- *Nejati, R. (2016). The durability of the effect of the frequent quizzes on Iranian high school students' vocabulary learning. *International Journal of the Humanities*, 23, 29–42. Retrieved from <http://journals.modares.ac.ir/article-27-6135-en.html>
- *Noll, V. H. (1939). The effect of written tests upon achievement in college classes: An experiment and a summary of evidence. *The Journal of Educational Research*, 32, 345–358. <http://dx.doi.org/10.1080/00220671.1939.10880843>
- *Norton, C. B. (2013). *The effect of frequent quizzing on student learning in a high school physical science classroom* (Master's thesis). Louisiana State University, Baton Rouge, LA. Retrieved from https://digitalcommons.lsu.edu/gradschool_theses/922/
- Nunes, L. D., & Weinstein, Y. (2012). Testing improves true recall and protects against the build-up of proactive interference without increasing false recall. *Memory*, 20, 138–154. <http://dx.doi.org/10.1080/09658211.2011.648198>
- *Nungester, R. J., & Duchastel, P. C. (1982). Testing versus review: Effects on retention. *Journal of Educational Psychology*, 74, 18–22. <http://dx.doi.org/10.1037/0022-0663.74.1.18>
- *Oglesby, R. (2013). *Examining nursing students' retention of taught content by repeat study and repeat testing: A replicated study* (Doctoral dissertation). Gardner-Webb University, Boiling Springs, NC. Retrieved from https://digitalcommons.gardner-webb.edu/nursing_etd/76
- *Olde Bekkink, M., Donders, R., van Muijen, G. N. P., & Ruiter, D. J. (2012). Challenging medical students with an interim assessment: A positive effect on formal examination score in a randomized controlled study. *Advances in Health Sciences Education*, 17, 27–37. <http://dx.doi.org/10.1007/s10459-011-9291-6>
- *Olsen, R. E., Weber, L. J., & Dörner, J. L. (1968). Quizzes as teaching aids. *Journal of Medical Education*, 43, 941–942.
- *Pagliarulo, C. L. (2011). *Testing effect and complex comprehension in a large introductory undergraduate biology course* (Doctoral dissertation). The University of Arizona, Tucson, AZ. Retrieved from <https://repository.arizona.edu/handle/10150/202773>
- Pan, S. C., Pashler, H., Potter, Z. E., & Rickard, T. C. (2015). Testing enhances learning across a range of episodic memory abilities. *Journal of Memory and Language*, 83, 53–61. <http://dx.doi.org/10.1016/j.jml.2015.04.001>
- Pan, S. C., & Rickard, T. C. (2018). Transfer of test-enhanced learning: Meta-analytic review and synthesis. *Psychological Bulletin*, 144, 710–756. <http://dx.doi.org/10.1037/bul0000151>
- Pastor, D. A., & Lazowski, R. A. (2018). On the multilevel nature of meta-analysis: A tutorial, comparison of software programs, and discussion of analytic choices. *Multivariate Behavioral Research*, 53, 74–89. <http://dx.doi.org/10.1080/00273171.2017.1365684>
- Pastötter, B., & Bäuml, K. H. (2014). Retrieval practice enhances new learning: The forward effect of testing. *Frontiers in Psychology*, 5, 286. <http://dx.doi.org/10.3389/fpsyg.2014.00286>
- Pastötter, B., & Bäuml, K. H. (2019). Testing enhances subsequent learning in older adults. *Psychology and Aging*, 34, 242–250. <http://dx.doi.org/10.1037/pag0000307>
- Pastötter, B., Schicker, S., Niedernhuber, J., & Bäuml, K. H. (2011). Retrieval during learning facilitates subsequent memory encoding. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37, 287–297. <http://dx.doi.org/10.1037/a0021801>
- Pastötter, B., Weber, J., & Bäuml, K. H. (2013). Using testing to improve learning after severe traumatic brain injury. *Neuropsychology*, 27, 280–285. <http://dx.doi.org/10.1037/a0031797>
- *Payne, B. C., & O'Malley, T. C. (2017). Blind pretesting and student performance in an undergraduate corporate finance course. *Journal of Financial Education*, 43, 47–62. Retrieved from <http://www.jstor.org/stable/90018418>
- *Pennebaker, J. W., Gosling, S. D., & Ferrell, J. D. (2013). Daily online testing in large classes: Boosting college performance while reducing achievement gaps. *PLoS ONE*, 8, e79774. <http://dx.doi.org/10.1371/journal.pone.0079774>

- *Petrović, J., Pale, P., & Jeren, B. (2017). Online formative assessments in a digital signal processing course: Effects of feedback type and content difficulty on students learning achievements. *Education and Information Technologies*, 22, 3047–3061. <http://dx.doi.org/10.1007/s10639-016-9571-0>
- Pham, B., Platt, R., McAuley, L., Klassen, T. P., & Moher, D. (2001). Is there a “best” way to detect and minimize publication bias?: An empirical evaluation. *Evaluation & the Health Professions*, 24, 109–125. <http://dx.doi.org/10.1177/016327870102400202>
- Phelps, R. P. (2012). The effect of testing on student achievement, 1910–2010. *International Journal of Testing*, 12, 21–43. <http://dx.doi.org/10.1080/15305058.2011.602920>
- Pierce, B. H., Gallo, D. A., & McCain, J. L. (2017). Reduced interference from memory testing: A postretrieval monitoring account. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 43, 1063–1072. <http://dx.doi.org/10.1037/xlm0000377>
- *Poljičanin, A., Carić, A., Vilović, K., Kosta, V., Marinović Guć, M., Aljinović, J., & Grković, I. (2009). Daily mini quizzes as means for improving student performance in anatomy course. *Croatian Medical Journal*, 50, 55–60. <http://dx.doi.org/10.3325/cmj.2009.50.55>
- *Purcell, J. B. (1974). *The effect of daily vocabulary quizzes on motivating pupils to study high school biology* (Doctoral dissertation). Loyola University Chicago, Chicago, IL. Retrieved from https://ecommons.luc.edu/luc_diss/1387
- Pustejovsky, J. E., & Rodgers, M. A. (2019). Testing for funnel plot asymmetry of standardized mean differences. *Research Synthesis Methods*, 10, 57–71. <http://dx.doi.org/10.1002/jrsm.1332>
- *Pyburn, D. T., Pazicni, S., Benassi, V. A., & Tappin, E. M. (2014). The testing effect: An intervention on behalf of low-skilled comprehenders in general chemistry. *Journal of Chemical Education*, 91, 2045–2057. <http://dx.doi.org/10.1021/ed4009045>
- Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language*, 60, 437–447. <http://dx.doi.org/10.1016/j.jml.2009.01.004>
- Rahman, S. (2017). *The effects of prequestions on classroom learning* (Master’s thesis). Iowa State University, IA. Retrieved from <https://lib.dr.iastate.edu/etd/15611>. <http://dx.doi.org/10.31274/etd-180810-5226>
- *Ramraje, S., & Sable, P. (2011). Comparison of the effect of post-instruction multiple-choice and short-answer tests on delayed retention learning. *The Australasian Medical Journal*, 4, 332–339. <http://dx.doi.org/10.4066/AMJ.2011.727>
- *Rashid, M. N., Soomro, A. M., Abro, A. H., & Noman, S. B. (2017). Medical students academic performance assessment in Physiology courses using formative and summative quizzes at SMBB Medical College Karachi, Pakistan. *Advances in Applied Physiology*, 2, 10–17. <http://dx.doi.org/10.11648/j.aap.20170201.12>
- *Raupach, T., Andresen, J. C., Meyer, K., Strobel, L., Koziolek, M., Jung, W., . . . Anders, S. (2016). Test-enhanced learning of clinical reasoning: A crossover randomised trial. *Medical Education*, 50, 711–720. <http://dx.doi.org/10.1111/medu.13069>
- *Rawson, K. A., Dunlosky, J., & Sciarrelli, S. M. (2013). The power of successive relearning: Improving performance on course exams and long-term retention. *Educational Psychology Review*, 25, 523–548. <http://dx.doi.org/10.1007/s10648-013-9240-4>
- *Rezaei, A. R. (2015). Frequent collaborative quiz taking and conceptual learning. *Active Learning in Higher Education*, 16, 187–196. <http://dx.doi.org/10.1177/1469787415589627>
- *Ritchie, S. J., Della Sala, S., & McIntosh, R. D. (2013). Retrieval practice, with or without mind mapping, boosts fact learning in primary school children. *PLoS ONE*, 8, e78976. <http://dx.doi.org/10.1371/journal.pone.0078976>
- *Rizi, A. R. B., & Tavakoli, M. (2015). The effects of the frequency of TOEFL iBT as quizzes on real-life reading comprehension tasks: The discourse in focus. *Journal of Applied Linguistics and Language Research*, 2, 80–92.
- *Robertson, W. L. (2010). *The impact of various quizzing patterns on the test performance of high school economics students* (Doctoral dissertation). Walden University, Minneapolis, MN. Retrieved from <https://search.proquest.com/openview/7bb5db240a0bf5ea616478a98adb4634/1?pq-origsite=gscholar&cbl=18750&diss=y>
- *Roediger, H. L., Agarwal, P. K., McDaniel, M. A., & McDermott, K. B. (2011). Test-enhanced learning in the classroom: Long-term improvements from quizzing. *Journal of Experimental Psychology: Applied*, 17, 382–395. <http://dx.doi.org/10.1037/a0026252>
- Roediger, H. L., & Karpicke, J. D. (2006a). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, 1, 181–210. <http://dx.doi.org/10.1111/j.1745-6916.2006.00012.x>
- Roediger, H. L., III, & Karpicke, J. D. (2006b). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, 17, 249–255. <http://dx.doi.org/10.1111/j.1467-9280.2006.01693.x>
- Roediger, H. L., III, Putnam, A. L., & Smith, M. A. (2011). Ten benefits of testing and their applications to educational practice. *Psychology of Learning and Motivation-Advances in Research and Theory*, 55, 1–36. <http://dx.doi.org/10.1016/B978-0-12-387691-1.00001-6>
- Roediger, H. L., & Marsh, E. J. (2005). The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31, 1155–1159. <http://dx.doi.org/10.1037/0278-7393.31.5.1155>
- *Ross, C. C., & Henry, L. K. (1939). The relation between frequency of testing and progress in learning psychology. *Journal of Educational Psychology*, 30, 604–611. <http://dx.doi.org/10.1037/h0055717>
- Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin*, 140, 1432–1463. <http://dx.doi.org/10.1037/a0037559>
- Rummer, R., Schweppe, J., Gerst, K., & Wagner, S. (2017). Is testing a more effective learning strategy than note-taking? *Journal of Experimental Psychology: Applied*, 23, 293–300. <http://dx.doi.org/10.1037/xap0000134>
- Ryan, R. M., & Deci, E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology*, 25, 54–67. <http://dx.doi.org/10.1006/ceps.1999.1020>
- *Sartain, A. F. (2018). *The frequency of testing and its effects on exam scores in a fundamental level baccalaureate nursing course* (Doctoral dissertation). University of Alabama Libraries, AL. Retrieved from <http://ir.ua.edu/handle/123456789/5405>
- *Schmidmaier, R., Ebersbach, R., Schiller, M., Hege, I., Holzer, M., & Fischer, M. R. (2011). Using electronic flashcards to promote learning in medical students: Retesting versus restudying. *Medical Education*, 45, 1101–1110. <http://dx.doi.org/10.1111/j.1365-2923.2011.04043.x>
- Schrank, Z. (2016). An assessment of student perceptions and responses to frequent low-stakes testing in introductory sociology classes. *Teaching Sociology*, 44, 118–127. <http://dx.doi.org/10.1177/0092055X15624745>
- Schunk, D. H. (1991). Self-efficacy and academic motivation. *Educational Psychologist*, 26, 207–231. http://dx.doi.org/10.1207/s15326985ep2603&4_2
- Schunk, D. H., Meece, J. R., & Pintrich, P. R. (2012). *Motivation in education: Theory, research, and applications* (4th ed.). Upper Saddle River, NJ: Pearson Higher Ed.
- Schwieren, J., Barenberg, J., & Dutke, S. (2017). The testing effect in the Psychology classroom: A meta-analytic perspective. *Psychology Learning & Teaching*, 16, 179–196. <http://dx.doi.org/10.1177/1475725717695149>
- *Sennhenn-Kirchner, S., Goerlich, Y., Kirchner, B., Notbohm, M., Schiekirka, S., Simmenroth, A., & Raupach, T. (2018). The effect of repeated testing vs repeated practice on skills learning in undergraduate

- dental education. *European Journal of Dental Education*, 22, e42–e47. <http://dx.doi.org/10.1111/eje.12254>
- *Sensenig, A. E. (2010). *Multiple choice testing and the retrieval hypothesis of the testing effect* (Doctoral dissertation). Colorado State University, CO. Retrieved from <https://psycnet.apa.org/record/2011-99080-461>
- *Shafiq, F., & Siddiquah, A. (2011). Effect of classroom quizzes on graduate students' achievement. *International Journal of Academic Research*, 3, 76–79.
- *Shapiro, A. M., & Gordon, L. T. (2012). A controlled study of clicker-assisted memory enhancement in college classrooms. *Applied Cognitive Psychology*, 26, 635–643. <http://dx.doi.org/10.1002/acp.2843>
- *Shirvani, H. (2009). Examining an assessment strategy on high school mathematics achievement: Daily quizzes Vs. weekly tests. *American Secondary Education*, 38, 34–45. Retrieved from <http://www.jstor.org/stable/41406065>
- *Silva, A. F. M. (2011). *The effect of evaluation on learning*. Porto, Portugal: Universidade Do Porto.
- Silver, N. C., & Dunlap, W. P. (1987). Averaging correlation coefficients: Should Fisher's z transformation be used? *Journal of Applied Psychology*, 72, 146–148. <http://dx.doi.org/10.1037/0021-9010.72.1.146>
- Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014a). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, 143, 534–547. <http://dx.doi.org/10.1037/a0033242>
- Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014b). P-curve and effect size: Correcting for publication bias using only significant results. *Perspectives on Psychological Science*, 9, 666–681. <http://dx.doi.org/10.1177/1745691614553988>
- Singapore Ministry of Education. (2018). "Learn for life" – preparing our students to excel beyond exam results [Press release]. Retrieved from <https://www.moe.gov.sg/news/press-releases/learn-for-life---preparing-our-students-to-excel-beyond-exam-results>
- Slamecka, N. J., & Katsaiti, L. T. (1988). Normal forgetting of verbal lists as a function of prior testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14, 716–727. <http://dx.doi.org/10.1037/0278-7393.14.4.716>
- *Slomer, A., & Chenkin, J. (2016). Does test-enhanced learning improve success rates of ultrasound-guided peripheral intravenous insertion? A randomized-controlled trial. *Canadian Journal of Emergency Medical Care*, 18, 310–315. <http://dx.doi.org/10.1017/cem.2016.296>
- Soderstrom, N. C., Kerr, T. K., & Bjork, R. A. (2016). The critical importance of retrieval—and spacing—for learning. *Psychological Science*, 27, 223–230. <http://dx.doi.org/10.1177/0956797615617778>
- *Son, J. Y., & Rivas, M. J. (2016). Designing clicker questions to stimulate transfer. *Scholarship of Teaching and Learning in Psychology*, 2, 193–207. <http://dx.doi.org/10.1037/std0000065>
- *Song, D., & Oh, E. Y. (2017). The effects of retrieval with different cues on second language vocabulary learning. *Journal of Language Learning and Teaching*, 7, 30–42. Retrieved from <http://dergipark.org.tr/jltl/issue/42178/507807>
- *Spreckelsen, C., & Juenger, J. (2017). Repeated testing improves achievement in a blended learning approach for risk competence training of medical students: Results of a randomized controlled trial. *BMC Medical Education*, 17, 177–186. <http://dx.doi.org/10.1186/s12909-017-1016-y>
- Stanley, T. D. (2017). Limitations of PET-PEESE and other meta-analysis methods. *Social Psychological and Personality Science*, 8, 581–591. <http://dx.doi.org/10.1177/1948550617693062>
- Stanley, T. D., & Doucouliagos, H. (2014). Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods*, 5, 60–78. <http://dx.doi.org/10.1002/jrsm.1095>
- Steele, D. (2011, January 25). Does test-taking help students learn? *New York Times* (Letter to the Editor), p. A24.
- *Steele, J. E. (2003). Effect of essay-style lecture quizzes on student performance on anatomy and physiology exams. *Bioscene*, 29, 15–20.
- *Steenhuis, H., Grinder, B., & Bruijn, E. J. (2009). The use(lessness) of online quizzes for achieving student learning. *International Journal of Information and Operations Management Education*, 3, 119–148. <http://dx.doi.org/10.1504/IJIOME.2009.031035>
- *Stenlund, T., Jönsson, F. U., & Jonsson, B. (2017). Group discussions and test-enhanced learning: Individual learning outcomes and personality characteristics. *Educational Psychology*, 37, 145–156. <http://dx.doi.org/10.1080/01443410.2016.1143087>
- Stenlund, T., Sundström, A., & Jonsson, B. (2016). Effects of repeated testing on short- and long-term memory performance across different test formats. *Educational Psychology*, 36, 1710–1727. <http://dx.doi.org/10.1080/01443410.2014.953037>
- *Strawitz, B. M. (1989). The effects of testing on science process skill achievement. *Journal of Research in Science Teaching*, 26, 659–664. <http://dx.doi.org/10.1002/tea.3660260802>
- Sullivan, D. (2017). Mediating test anxiety through the testing effect in asynchronous, objective, online assessments at the university level. *Journal of Education and Training*, 4, 107–123. <http://dx.doi.org/10.5296/jet.v4i2.10777>
- Szpunar, K. K., Jing, H. G., & Schacter, D. L. (2014). Overcoming overconfidence in learning from video-recorded lectures: Implications of interpolated testing for online education. *Journal of Applied Research in Memory & Cognition*, 3, 161–164. <http://dx.doi.org/10.1016/j.jarmac.2014.02.001>
- Szpunar, K. K., Khan, N. Y., & Schacter, D. L. (2013). Interpolated memory tests reduce mind wandering and improve learning of online lectures. *Proceedings of the National Academy of Sciences*, 110, 6313–6317. <http://dx.doi.org/10.1073/pnas.1221764110>
- Szpunar, K. K., McDermott, K. B., & Roediger, H. L. (2007). Expectation of a final cumulative test enhances long-term retention. *Memory & Cognition*, 35, 1007–1013. <http://dx.doi.org/10.3758/BF03193473>
- Szpunar, K. K., McDermott, K. B., & Roediger, H. L. (2008). Testing during study insulates against the buildup of proactive interference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34, 1392–1399. <http://dx.doi.org/10.1037/a0013082>
- *Tepeç, M., & Çevik, Y. D. (2018). Comparison of three instructional strategies in teaching programming: Restudying material, testing and worked example. *Journal of Learning and Teaching in Digital Age*, 3, 42–50.
- *Teplitski, M., Irani, T., Krediet, C. J., Di Cesare, M., & Marvasi, M. (2018). Student-generated pre-exam questions is an effective tool for participatory learning: A case study from ecology of waterborne pathogens course. *Research in Food Science Education*, 17, 76–84. <http://dx.doi.org/10.1111/1541-4329.12129>
- *Terenyi, J., Anksorus, H., & Persky, A. M. (2018). Impact of spacing of practice on learning brand name and generic drugs. *American Journal of Pharmaceutical Education*, 82, 6179. <http://dx.doi.org/10.5688/ajpe6179>
- Terrin, N., Schmid, C. H., Lau, J., & Olkin, I. (2003). Adjusting for publication bias in the presence of heterogeneity. *Statistics in Medicine*, 22, 2113–2126. <http://dx.doi.org/10.1002/sim.1461>
- Thomas, R. C., & McDaniel, M. A. (2013). Testing and feedback effects on front-end control over later retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 39, 437–450. <http://dx.doi.org/10.1037/a0028886>
- Thompson, C. P., Wenger, S. K., & Bartling, C. A. (1978). How recall facilitates subsequent recall: A reappraisal. *Journal of Experimental Psychology: Human Learning and Memory*, 4, 210–221. <http://dx.doi.org/10.1037/0278-7393.4.3.210>
- Thompson, V. A., & Campbell, J. I. D. (2004). A power struggle: Between- vs. within-subjects designs in deductive reasoning research. *Psychologia*, 47, 277–296. <http://dx.doi.org/10.2117/psysoc.2004.277>

- Tobias, S. (1985). Test anxiety: Interference, defective skills, and cognitive capacity. *Educational Psychologist*, 20, 135–142. http://dx.doi.org/10.1207/s15326985ep2003_3
- *Trumbo, M. C., Leiting, K. A., McDaniel, M. A., & Hodge, G. K. (2016). Effects of reinforcement on test-enhanced learning in a large, diverse introductory college psychology course. *Journal of Experimental Psychology: Applied*, 22, 148–160. <http://dx.doi.org/10.1037/xap0000082>
- Tse, C. S., & Pu, X. (2012). The effectiveness of test-enhanced learning depends on trait test anxiety and working-memory capacity. *Journal of Experimental Psychology: Applied*, 18, 253–264. <http://dx.doi.org/10.1037/a0029190>
- *Tuckman, B. W. (2000, November). *Using frequent testing to increase students' motivation to achieve*. Paper presented at the 7th Biannual International Conference on Motivation, Leuven, Belgium.
- Van Den Noortgate, W., & Onghena, P. (2003). Multilevel meta-analysis: A comparison with traditional meta-analytical procedures. *Educational and Psychological Measurement*, 63, 765–790. <http://dx.doi.org/10.1177/0013164403251027>
- *Van Deventer, I. A. (2014). *The effects of different types of frequent summative testing on student achievement in first-year financial accounting courses* (Doctoral dissertation). Northcentral University, San Diego, CA. Retrieved from <http://libproxy1.nus.edu.sg/login?url=https://search-proquest-com.libproxy1.nus.edu.sg/docview/1561161405?accountid=13876>
- *Van Deventer, I. A. (2015). The use of traditional summative testing to maximize student achievement: An empirical study. *Journal of Higher Education Theory and Practice*, 15, 61–66.
- Vaughn, K. E., & Rawson, K. A. (2011). Diagnosing criterion-level effects on memory: What aspects of memory are enhanced by repeated retrieval? *Psychological Science*, 22, 1127–1131. <http://dx.doi.org/10.1177/0956797611417724>
- Veltre, M. T., Cho, K. W., & Neely, J. H. (2015). Transfer-appropriate processing in the testing effect. *Memory*, 23, 1229–1237. <http://dx.doi.org/10.1080/09658211.2014.970196>
- Vevea, J. L., & Hedges, L. V. (1995). A general linear model for estimating effect size in the presence of publication bias. *Psychometrika*, 60, 419–435. <http://dx.doi.org/10.1007/BF02294384>
- Vevea, J. L., & Woods, C. M. (2005). Publication bias in research synthesis: Sensitivity analysis using a priori weight functions. *Psychological Methods*, 10, 428–443. <http://dx.doi.org/10.1037/1082-989X.10.4.428>
- Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, 36, 1–48. <http://dx.doi.org/10.18637/jss.v036.i03>
- *Vojdanoska, M., Cranney, J., & Newell, B. R. (2010). The testing effect: The role of feedback and collaboration in a tertiary classroom setting. *Applied Cognitive Psychology*, 24, 1183–1195. <http://dx.doi.org/10.1002/acp.1630>
- *Voss, J. (2014). *Effect of retrieval practice on applied knowledge: Evidence from a professional training program* (Doctoral dissertation). Washington University in St. Louis, St. Louis, MO. Retrieved from https://openscholarship.wustl.edu/art_sci_etds/327/
- Wang, B., & Zhao, C. (2019). Testing the retrieval effort theory. *Swiss Journal of Psychology*, 78, 125–136. <http://dx.doi.org/10.1024/1421-0185/a000229>
- Weinstein, Y., Gilmore, A. W., Szpunar, K. K., & McDermott, K. B. (2014). The role of test expectancy in the build-up of proactive interference in long-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 40, 1039–1048. <http://dx.doi.org/10.1037/a0036164>
- Weinstein, Y., McDermott, K. B., & Szpunar, K. K. (2011). Testing protects against proactive interference in face-name learning. *Psychonomic Bulletin & Review*, 18, 518–523. <http://dx.doi.org/10.3758/s13423-011-0085-x>
- Westrick, P. A. (2017). Reliability estimates for undergraduate grade point average. *Educational Assessment*, 22, 231–252. <http://dx.doi.org/10.1080/10627197.2017.1381554>
- Wheeler, M. A., & Roediger, H. L., III. (1992). Disparate effects of repeated testing: Reconciling Ballard's (1913) and Bartlett's (1932) results. *Psychological Science*, 3, 240–246. <http://dx.doi.org/10.1111/j.1467-9280.1992.tb00036.x>
- *Wickline, V. B., & Spektor, V. G. (2011). Practice (rather than graded) quizzes, with answers, may increase introductory psychology exam performance. *Teaching of Psychology*, 38, 98–101. <http://dx.doi.org/10.1177/0098628311401580>
- Wigfield, A., & Eccles, J. S. (2000). Expectancy–value theory of achievement motivation. *Contemporary Educational Psychology*, 25, 68–81. <http://dx.doi.org/10.1006/ceps.1999.1015>
- *Wiklund-Hörnqvist, C., Jonsson, B., & Nyberg, L. (2014). Strengthening concept learning by repeated testing. *Scandinavian Journal of Psychology*, 55, 10–16. <http://dx.doi.org/10.1111/sjop.12093>
- Wissman, K. T., Rawson, K. A., & Pyc, M. A. (2011). The interim test effect: Testing prior material can facilitate the learning of new material. *Psychonomic Bulletin & Review*, 18, 1140–1147. <http://dx.doi.org/10.3758/s13423-011-0140-7>
- *Yalaki, Y. (2010). Simple formative assessment, high learning gains in college general chemistry. *Eurasian Journal of Educational Research*, 40, 17–31.
- *Yalaki, Y., & Bayram, Z. (2015). Effect of formative quizzes on teacher candidates' learning in general chemistry. *International Journal of Research in Education*, 1, 151–156. http://yunus.hacettepe.edu.tr/~yyalaki/yayinlar/Yalaki-Bayram_IJRES.pdf
- *Yamin, S. B. (1988). *Frequency of testing and its effects on achievement, test anxiety and attitudes toward science of students at University Technology of Malaysia* (Doctoral dissertation). Oregon State University, Corvallis, OR. Retrieved from <http://hdl.handle.net/1957/28982>
- Yan, V. X., Thai, K.-P., & Bjork, R. A. (2014). Habits and beliefs that guide self-regulated learning: Do they vary with mindset? *Journal of Applied Research in Memory & Cognition*, 3, 140–152. <http://dx.doi.org/10.1016/j.jarmac.2014.04.003>
- Yang, C., Chew, S.-J., Sun, B., & Shanks, D. R. (2019). The forward effects of testing transfer to different domains of learning. *Journal of Educational Psychology*, 111, 809–826. <http://dx.doi.org/10.1037/edu0000320>
- Yang, C., Potts, R., & Shanks, D. R. (2017). The forward testing effect on self-regulated study time allocation and metamemory monitoring. *Journal of Experimental Psychology: Applied*, 23, 263–277. <http://dx.doi.org/10.1037/xap0000122>
- Yang, C., Potts, R., & Shanks, D. R. (2018). Enhancing learning and retrieval of new information: A review of the forward testing effect. *npj Science of Learning*, 3, 8. <http://dx.doi.org/10.1038/s41539-018-0024-y>
- Yang, C., & Shanks, D. R. (2018). The forward testing effect: Interim testing enhances inductive learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 44, 485–492. <http://dx.doi.org/10.1037/xlm0000449>
- Yang, C., Sun, B., Potts, R., Yu, R., & Shanks, D. R. (2020). Do working memory capacity and test anxiety modulate the beneficial effects of testing on new learning? *Journal of Experimental Psychology: Applied*. Advance online publication. <http://dx.doi.org/10.1037/xap0000278>
- Yeager, D. S., Hanselman, P., Walton, G. M., Murray, J. S., Crosnoe, R., Muller, C., . . . Dweck, C. S. (2019). A national experiment reveals where a growth mindset improves achievement. *Nature*, 573, 364–369. <http://dx.doi.org/10.1038/s41586-019-1466-y>
- Yue, C. L., Soderstrom, N. C., & Bjork, E. L. (2015). Partial testing can potentiate learning of tested and untested material from multimedia lessons. *Journal of Educational Psychology*, 107, 991–1005. <http://dx.doi.org/10.1037/edu0000031>

- *Zamini, G., Khadem Erfan, M. B., Rahmani, M. R., Khodavaisy, M. S., & Davari, B. (2013). Effects of frequent announced parasitology quizzes on the academic achievement. *Iranian Journal of Parasitology*, 8, 617–621.
- *Zarei, A. A. (2015). On the effectiveness of quizzes on L2 idioms learning. *Iranian Journal of Language Testing*, 5, 60–77.
- *Zayac, R. M., Ratkos, T., Frieder, J. E., & Paulk, A. (2016). A comparison of active student responding modalities in a general psychology course. *Teaching of Psychology*, 43, 43–47. <http://dx.doi.org/10.1177/0098628315620879>
- *Zhang, N., & Henderson, C. N. R. (2015). Can formative quizzes predict or improve summative exam performance? *The Journal of Chiropractic Education*, 29, 16–21. <http://dx.doi.org/10.7899/JCE-14-12>
- Zheng, M., & Bender, D. (2019). Evaluating outcomes of computer-based classroom testing: Student acceptance and impact on learning and exam performance. *Medical Teacher*, 41, 75–82. <http://dx.doi.org/10.1080/0142159X.2018.1441984>
- Zhou, A., Yang, T., Cheng, C., Ma, X., & Zhao, J. (2015). Retrieval practice produces more learning in multiple-list tests with higher-order skills. *Acta Psychologica Sinica*, 47, 928. <http://dx.doi.org/10.3724/SP.J.1041.2015.00928>
- *Zhu, C., & Urhahne, D. (2018). The use of learner response systems in the classroom enhances teachers' judgment accuracy. *Learning and Instruction*, 58, 255–262. <http://dx.doi.org/10.1016/j.learninstruc.2018.07.011>

Received December 3, 2019

Revision received June 13, 2020

Accepted September 1, 2020 ■