Retrieval Practice Consistently Benefits Student Learning: A Systematic Review of Applied Research in Schools and Classrooms

Pooja K. Agarwal, Berklee College of Music (ORCID 0000-0001-8880-8650)

Ludmila D. Nunes, Association for Psychological Science (ORCID 0000-0003-4467-4341)

Janell R. Blunt, Anderson University (ORCID 0000-0002-1660-084X)

Accepted, *Educational Psychology Review* (2023)

Author Note

We would like to thank Henry L. Roediger, III and Doug Rohrer for valuable discussions, as well four anonymous reviewers for their feedback. We also thank Tarah Collins, Jessica Bailey, Erin Eberly, Makayla Miller, Emily Glassman, and Lauren Stephen for their assistance.

This project is registered on the Open Science Framework (OSF) at http://osf.io/mz2ks.

Correspondence should be addressed to Pooja K. Agarwal, Ph.D. at pagarwal@berklee.edu.

Abstract

Given the growing interest in retrieval practice among educators, it is valuable to know when retrieval practice does and does not improve student learning—particularly for educators who have limited classroom time and resources. In this literature review, we developed a narrow operational definition for "classroom research" compared to previous reviews of the literature. We screened nearly 2,000 abstracts and systematically coded 50 experiments to establish a clearer picture of benefits from retrieval practice in real world educational settings. Our review yielded 49 effect sizes and a total n = 5,374, the majority of which (57%) revealed medium or large benefits from retrieval practice. We found that retrieval practice improved learning for a variety of education levels, content areas, experimental designs, final test delays, retrieval and final test formats, and timing of retrieval practice and feedback; however, only 6% of experiments were conducted in non-WEIRD countries. Based on our review of the literature, we make eight recommendations for future research and provide educators with a better understanding of the robust benefits of retrieval practice across a range of school and classroom settings.

Keywords: Retrieval practice, learning, classroom, schools, applied research, testing effect

Declarations

Funding: Not applicable

Conflicts of interest/Competing interests: None

Availability of data and material: All data are available at http://osf.io/mz2ks

Code availability: Syntax used for the literature search is listed in Table 1 in the manuscript

Authors' contributions: Pooja Agarwal developed the idea for this literature review. Pooja Agarwal and Ludmila Nunes developed the methodology and coding procedure. Ludmila Nunes developed the search syntax and executed the search of databases. All authors screened abstracts and Ludmila Nunes downloaded all articles that passed detailed screening. All authors coded the articles and Ludmila Nunes calculated effect sizes when needed. Pooja Agarwal drafted the manuscript, which was edited by Ludmila Nunes and Janell Blunt.

Additional notes: This research is registered on the Open Science Framework (doi:10.17605/osf.io/mz2ks) and publicly available at http://osf.io/mz2ks. Because our manuscript is a literature review, human subjects approval was not required. We declare that there are no conflicts of interest with respect to the authorship or the publication of this article, and all authors have approved the submitted manuscript. This research is not under consideration at other journals.

Retrieval Practice Consistently Benefits Student Learning:
A Systematic Review of Applied Research in Schools and Classrooms

As researchers have demonstrated for more than a century, retrieval practice—the act of recalling previously learned information—improves long-term learning and memory. For example, simply retrieving a trivia fact (e.g., what was the name of the ship of Charles Darwin's famous voyage?) helps students remember this fact better than if they simply re-read it multiple times.[1]

In a landmark study on retrieval practice by Roediger and Karpicke (2006b), college students read brief passages (e.g., about sea otters, the solar system, etc.) and either engaged in re-reading the passages or retrieval practice (i.e., free recall by writing down everything they could remember from the passage). After a five-minute delay, students performed better on a free recall test after initially re-reading, but after a one-week delay, their performance was greater when they engaged in initial retrieval practice by freely recalling the passage.

Early research on the use of retrieval practice as a strategy to improve long-term learning found consistent benefits from retrieval practice (e.g., Gates, 1917; Glover, 1989; Myers, 1914; Spitzer, 1939). Revived by Roediger and Karpicke (2006b), this area of research has been progressing rapidly. In particular, scientists have been extending research from laboratory settings to educational settings, with demonstrated benefits for student learning ranging from primary school to medical school, and content areas ranging from math and science to history and foreign languages (Dunlosky et al., 2013). Researchers are increasingly urging educators to apply retrieval practice in the classroom (Agarwal & Bain, 2019; Brown, Roediger, & McDaniel, 2014; Butler & Carpenter, 2015; Dunlosky & Rawson, 2019; Fazio & Marsh, 2019; Karpicke,

[1] The name of the ship of Charles Darwin's famous voyage is the Beagle.

Blunt, & Smith, 2016; Nunes & Karpicke, 2015; Roediger & Butler, 2011; Weinstein, Madan, & Sumeracki, 2018). In tandem, educators are increasingly implementing retrieval practice using a variety of methods including multiple-choice online apps, frequent recall prompts, and quizzes during lectures.

Given the growing interest in retrieval practice among educators, it is valuable to know when retrieval practice does and does not improve student learning—particularly for educators who have limited classroom time and resources. From a scientific standpoint, it is also valuable to have a clear understanding of the literature to date in order to inform future research.

To address these pursuits for both educators and researchers, we developed a narrow operational definition for "classroom research" compared to previous reviews of the literature. Using our definition and detailed search syntax (see Table 1), we screened nearly 2,000 studies and coded 50 experiments drawn from research on retrieval practice, conducted in classroom settings.

Aims of the Present Study

Aim 1: Compare Apples to Apples Using Narrower Review Criteria

The first aim of the present review was to compare classroom studies that examined the effects of retrieval practice, while applying a narrower set of inclusion criteria than used in prior reviews or meta-analyses. In order to inform future research (Aim 2) and clarify recommendations for educators (Aim 3), we included only classroom studies in which retrieval practice was administered individually and in person; in other words, we did not include studies conducted in laboratory settings, studies with collaborative retrieval, nor studies in which retrieval practice was administered online.

To date, reviews and meta-analyses of research on retrieval practice have typically

included a mix of studies drawn from both laboratory and applied settings (Adesope et al., 2017; Bangert-Drowns, Kulik, & Kulik, 1991; Brame & Biel, 2015; Dunlosky et al., 2013; Eisenkraemer, Jaeger, & Stein, 2013; Green, Moeller, & Spak, 2018; Karpicke & Grimaldi, 2012; McLaughlin & Coderre, 2015; Nguyen & McDaniel, 2015; Pyc, Agarwal, & Roediger, 2012; Rowland, 2014). For example, Adesope et al. (2017) conducted a meta-analysis of 217 research studies on retrieval practice, of which 11% were from classroom settings (p. 666). While informative for research purposes, what works in the laboratory does not necessarily work in the classroom—and vice versa. As Adesope et al. observed in their own meta-analysis, "In light of potential confounds, comparison of classroom and laboratory effect sizes should be interpreted with caution" (p. 687).

Even when meta-analyses and reviews were restricted to research in applied settings, the format and implementation of retrieval practice for included studies varied widely. For example, in the 23 studies included in the literature review by Moreira et al. (2019), materials ranged from encyclopedia passages (Jaeger et al., 2015) to TV recordings (Cranney et al., 2009). One study included in the Moreira et al. review was conducted in the classroom, but students worked individually at computer stations (Lipko-Speed, Dunlosky, & Rawson, 2014). Another study was carried out after the course had concluded (Carpenter, Pashler, & Cepeda, 2009).

Of the studies included in meta-analyses by Schwieren, Barenberg, and Dutke (2017; 19 studies) and Sotola and Crede (2020; 52 studies), students engaged in unsupervised online retrieval practice (Burdo & O'Dwyer, 2015; Daniel & Broida, 2004; Kibble, 2007), collaborative retrieval practice (Bojnova & Oigara, 2011; Vojdanoska, Cranney, & Newell, 2010), and retrieval practice in computer labs after the classroom lecture (Wiklund-Hornqvist et al., 2014).

While prior reviews and meta-analyses have increased our overall understanding of the

benefits of retrieval practice across a variety of educationally-relevant materials and conditions, recommendations drawn from such widely varying circumstances may lack the specificity required for actual classroom practice. As an alternative, we present a comprehensive database search capturing more published articles than previous reviews. We also used narrow screening criteria to ensure that the classroom studies manipulated retrieval practice under comparable real world conditions.

As an additional cause for concern regarding the existing literature, recent works have pointed out several shortcomings of meta-analytical methods, including the use of random effects models, when data to be analyzed are complex (e.g., Carter, Schönbrodt, Gervais, & Hilgard, 2019; McShane & Böckenholdt, 2020; Tipton, Pustejovsky, & Ahmadi, 2019). Applied research is inherently complex and messy, with challenges and confounds that are difficult to control. For example, some of the studies in the present review included multiple retrieval practice conditions compared against the same control condition, one retrieval practice condition compared against multiple control conditions, or data that were collapsed across conditions—all of which render the effect sizes computed non-independent. In addition, research has shown that the reproducibility of mean effect sizes derived from meta-analyses is low (Lakens, Hilgard, & Staaks, 2016).

Furthermore, reporting of data has been inconsistent in prior reviews. For example, Adesope et al. (2017) and Rowland (2014) reported mean weighted effect sizes, but they did not report effect sizes or confidence intervals for individual studies. Similarly, Brame and Biel (2015), Green et al. (2018), and Moreira et al. (2019) did not report effect sizes for individual studies. Bangert-Drowns et al. (1991) reported effect sizes for individual studies, but they did not report sample sizes nor confidence intervals around the effect sizes.

Thus, reviews and meta-analyses in this area of applied research should be considered with caution. In consideration of these challenges—a mix of settings, varied formats of retrieval practice, wide-ranging implementation methods, and statistical concerns—we felt that conditions across the 50 experiments included in the present review were too varied for a meta-analytic approach. Even under our narrowed inclusion criteria, seldom was there a consistent retrieval practice format or implementation method, a group of subjects without attrition, a perfectly controlled experimental condition or dependent measure, or a singular effect of interest.

For these reasons, in lieu of a meta-analysis that collapses effect sizes over a wide variety of applied experiments, we report effect sizes for each individual comparison in forest plots (Figures 3-6) and also in the Appendix. Because we include sample sizes, effect sizes, and confidence intervals for each individual study and comparison (when data were available to calculate them), the present review provides a more accurate understanding of the conditions under which retrieval practice benefits learning compared to prior reviews.

Our operational definition of retrieval practice was as follows: an active attempt by a student to recall or recognize, and then reconstruct, their memory of knowledge during initial learning. While this is sometimes referred to as the "testing effect," we chose to use the phrase "retrieval practice" in our review of classroom research for a few reasons.

First, the term "retrieval practice" has become more commonly used in the research literature to encompass various forms of retrieval during initial learning, including both recall and recognition (e.g., Karpicke, 2012).

Second, retrieval practice in the classroom takes many forms that differ from the typical notion of a test. For example, across the 50 experiments included in our review, students engaged in a variety of retrieval practice activities, including free recall, short answer quizzes, multiple-

choice quizzes, and quizzes with standardized patients. These low-stakes or no-stakes classroom learning activities were seldom referred to as "tests" by the authors of the studies. In addition, consider the increasing use of educational apps for retrieval practice, such as Kahoot and Quizlet, which bear very little resemblance to traditional tests.

Third, the terms "testing," "testing effect," and "test-enhanced learning" create confusion with unrelated educational activities such as summative assessments and standardized testing (Agarwal & Bain, 2019). As such, our operational definition and terminology highlight that it is the *process* of practicing retrieval (the active attempt) that shapes learning, not tests.

After screening nearly 2,000 abstracts, 50 experiments drawn from 37 studies met our full screening criteria, in which we required classroom-relevant materials, retrieval practice by students individually, and implementation during class periods under the supervision of the instructor or researcher. Critically, we compared classroom studies only (apples to apples), rather than drawing comparisons across both classroom and laboratory settings (apples to oranges). We feel that our narrowed criteria provides greater specificity in terms of directions for future research (Aim 2) and recommendations for classroom implementation (Aim 3).

Aim 2: Inform Future Directions for Research on Retrieval Practice

A second aim of the present review was to inform future directions for research on retrieval practice. In order to construct the most thorough review of the literature possible, we developed precise search syntax for five databases: PsycINFO, PsycARTICLES, ERIC, Web of Science, and PubMed (see Table 1). While our screening criteria were narrower compared to previous reviews, our systematic search yielded more classroom-specific peer-reviewed publications (37 studies in the present review of the literature) when compared to the number of studies included in previous reviews (e.g., 30 studies in Adesope et al., 2017; 23 studies in

Moreira et al., 2019; 19 studies in Schwieren et al., 2017)

It is possible that the greater number of studies in our review, compared to previous reviews, may be due to a more recent search of the literature. Still, using our methodology, we found a few classroom studies that were not included in previous literature reviews, albeit having been published before those reviews (e.g., Graham, 1999; Kromann et al., 2009; Narloch et al., 2006). In this way, we feel that our systematic review contributes a comprehensive record of research to date, which better informs future directions for applied research on retrieval practice.

Specifically, we investigated unresolved questions in the research literature on retrieval practice. For example, is there an optimal frequency of retrieval practice to improve student learning in classroom settings? Do all content areas and educational levels (e.g., K-12, undergraduate, and medical school) benefit from retrieval practice? Which is more beneficial for student learning, multiple-choice or short answer retrieval practice? To foreshadow our results, we examined these possible moderating variables by categorizing the 50 included experiments across various characteristics (e.g., education level, experimental design, sample size, etc.). We present effect sizes for each individual experiment in forest plots (Figures 3-6) and also in the Appendix.

Typically, we think of laboratory research as informing applications of cognitive psychology in real world settings. The opposite is also true: Research conducted in real world settings can inform basic research in laboratory settings. For example, are benefits from retrieval practice modulated by incentives? Researchers have examined this question in laboratory experiments with foreign language vocabulary, but have found inconsistent results (e.g., Abel & Bäuml, 2020; Kang & Pashler, 2014). Meanwhile, in the real world, students' performance on retrieval practice and final tests typically count toward course grades; in most of the studies

included in the present review (66%), this was the case. In contrast to previous reviews of retrieval practice research in classrooms (e.g., Sotola & Crede, 2020), we intentionally included grades as a moderating variable in our coding system. With a better understanding of motivational factors in the classroom in the present review, researchers can more effectively examine these factors in the laboratory.

Aim 3: Clarify Recommendations for Classroom Implementation of Retrieval Practice

In order to provide recommendations for classroom implementation of retrieval practice, we systematically coded conditions of interest to educators (see the Appendix and http://osf.io/mz2ks for the complete coding). For example, educators express a number of concerns regarding implementation, particularly the use of multiple-choice questions and the optimal timing for feedback (Agarwal & Bain, 2019). In both of these examples, laboratory and applied research suggest mixed approaches (Adesope et al., 2017). Additional conditions of interest in the present review included:

- Education level (e.g., K-12, college/university, medical school)
- Content area (e.g., psychology, medicine, and history)
- Comparison conditions (e.g., reviewing material, lessons without quizzes, infrequent high stakes exams)
- Retrieval practice timing (e.g., every class, once a week, once a month)
- Length of delay between the last retrieval opportunity and the final test (e.g., days or weeks)
- Format for initial retrieval practice and final test (e.g., multiple-choice, short answer, free recall)
- Feedback timing (e.g., immediate, delayed, no feedback)

Our narrower review criteria also allowed us to examine the benefits of retrieval practice under conditions found in real world classrooms—but uncommon in laboratory studies—including research with diverse student populations and situations in which performance on retrieval practice counted toward students' course grades.

Considerations for Applied Research in Classrooms

Definition of "Classroom" Research

The primary aim of the present review was to examine the literature on retrieval practice research conducted in classrooms. But how does one define a "classroom?" Classrooms, particularly in the present day, take a variety of forms (e.g., small seminars, large lecture halls, and online) and include a variety of activities (e.g., lectures, group projects, and discussion). For example, 14% of all students in higher education (more than 5.8 million students) take the entirety of their courses online (Allen & Seaman, 2016).

In addition, retrieval practice is inherent to classroom instruction: teachers pose questions during class, students retrieve knowledge during group discussion, and students retrieve during course exams. If we were to consider any classroom settings in which retrieval practice takes place, such a review would be far too broad to draw conclusions for future research and practical implementation. Thus, defining what constitutes a classroom for the purpose of this review required careful consideration.

Consider a study conducted by Herbert Spitzer in 1939. More than 3,500 children across the state of Iowa were asked to read passages about peanuts and bamboo, which were followed by zero, one, two, or three multiple-choice tests administered in the classroom. After two months, final test performance was greater for students who engaged in retrieval practice compared to students who did not receive initial retrieval practice.

While this study by Spitzer (1939) was conducted with school-age children in classrooms, Spitzer himself observed, “The learning was of little practical use to the children” (p. 655). This study is a valuable demonstration of the benefits of retrieval practice in an applied setting, but it is also an instance in which educational research deviated from typical classroom instruction; namely, the materials were irrelevant to what students were learning in class (see also Myers, 1914).

As a second example, consider a study by Sennhenn-Kirchner et al. (2016). Dental students completed a four-hour course on suturing skills, which was followed by retrieval practice using a suture simulation pad, either collaboratively in pairs or individually. On a final test approximately one month later, students in the collaborative retrieval group outperformed students who had engaged in retrieval practice individually.

In Sennhenn-Kirchner et al. (2016), the information to be learned and the format of retrieval practice was typical of dental education. However, in the collaborative retrieval condition, it is possible that one student led suture practice while the other student watched and did not engage in retrieval practice. The extent to which all students engaged in retrieval practice was not measured and could not be guaranteed by the researchers or instructors.

In order to ascertain trends and draw conclusions from the growing literature on retrieval practice, we limited our definition of “classroom” research using the following guidelines:

- Relevant course materials: Information to be learned for research purposes was the same as, or directly related to, assigned course materials
- Individual, not collaborative: All students engaged in retrieval practice individually under the supervision of researchers and instructors
- Closed-book, not open-book: All retrieval practice took place without the use of notes,

external learning aids, or the internet

Comparison Conditions for Experiments in Classroom Settings

When it comes to control conditions in research on learning, it is critical to compare *what* students do and also for how *long*. In laboratory experiments, retrieval practice is typically compared to re-studying, particularly to ensure that time spent with materials is equated across conditions (e.g., Roediger & Karpicke, 2006b). In classroom experiments, the comparison to retrieval practice is typically "business as usual," where a teacher lectures on the same material, but without quizzes (Khanna & Cortese, 2016).

For all studies reviewed in the present study, students spent approximately the same amount of time with materials between the retrieval practice intervention and comparison conditions. For example, in Freda and Lipp (2016), classes consisted of lectures with quizzes vs. lectures without quizzes. In Kromann et al. (2010), students received quizzes on simulated cardiac arrest scenarios or they received lecture presentations of the scenarios. In Roediger, Agarwal, McDaniel, and McDermott (2011), using a within-student design, questions over half the material were presented on quizzes and the final exam, whereas questions on the remaining material appeared only on the final exam (non-quizzed items).

In the present review, we chose to include two studies in which the comparison of interest was *dosage*, where researchers directly manipulated quantity, or the number of opportunities for retrieval practice (Dunlosky et al., 2013). We included experiments by Foss and Pirozzolo (2017), in which retrieval practice in the form of 4-8 exams was compared to learning after 2-3 exams. We also included an experiment by Leeming (2002), in which quizzes administered during every class were compared to four exams over the course of the semester. All remaining studies included in the present review (35 studies) did not directly manipulate dosage and many

did not report dosage of retrieval practice at all.

Intertwined with dosage of retrieval practice are timing and spacing, which presents a few challenges—especially in classroom research. First, different dosages of retrieval practice imply different retrieval practice timings (e.g., every class vs. every week represents a higher dosage but also a different implementation schedule). In this way, dosage can be easily confounded with spacing (e.g., quizzes every class means more retrieval practice, but less spacing than once per week). Second, higher dosage should result in more learning, but we also know that increased spacing can result in more learning (McDaniel, Agarwal, Huelser, McDermott, & Roediger, 2011). Third, many of the studies included in our review confounded dosage, timing, and spacing, or they did not report whether repeated retrieval opportunities included the same items (a requirement for spaced practice). Thus, because we could not code for dosage or spacing, we coded for *retrieval practice timing* across the studies (e.g., whether students generally engaged in retrieval practice daily, weekly, monthly, etc.).

We chose to exclude studies in which the primary comparison of interest was between two (or more) *types* of retrieval practice. For example, we excluded experiments by Carpenter et al. (2016; they compared retrieval practice via drawing diagrams vs. labeling diagrams), Niedermeyer and Sullivan (1972; they compared three vs. four multiple-choice test alternatives), Rohrer, Dedrick, Hartwig, and Cheung (2020; they compared blocked vs. interleaved retrieval practice), and Weinstein, Nunes, and Karpicke (2016, Experiment 3; they compared quiz questions interspersed during lectures or at the end of lectures). This growing area of applied research on optimal types of retrieval practice awaits a future review of the literature.

Note that in the present review, we do not refer to “control conditions” or “control groups.” Instead, we refer to “*comparison* conditions” and “comparison groups.” What may be

considered an appropriate control group in one classroom setting may not be appropriate in another, thus we did not restrict whether a control had to take a specific form of re-studying, re-presentation, repetition, concept mapping, or alternative "non-retrieval" conditions (see also Kornell, Rabelo, & Klein, 2012).

Internal and External Validity in Applied Research

Applied research, particularly on student learning and memory, brings with it a number of uncontrolled variables and circumstances. These variables for both students and teachers—such as absences, external motivators, and commitments outside the classroom—can affect internal validity and fidelity of implementation, or the extent to which an intervention is implemented in accordance with the procedure (O'Donnell, 2008).

A common characterization of laboratory and classroom research is that laboratory research is regarded as high on internal validity (free from errors in the experiment) and low on external validity (findings do not generalize to the real world), while classroom research is low on internal validity and high on external validity. Even so, as Anderson, Lindsay, and Bushman (1999) argue, *individual* studies may be high or low on internal and external validity; validity cannot be defined simply based on whether a study was conducted in a lab or a classroom.

In addition, in any situation in which retrieval practice was implemented without supervision, it is impossible to know whether it conformed to our operational definition. Thus, in order to maintain fidelity of implementation and internal validity as much as possible, our inclusion criteria required that all instructional activities, retrieval practice, and assessments took place in person under the supervision of the researcher or the instructor, in person, and not online.

By screening nearly 2,000 abstracts and systematically reviewing 50 experiments

conducted in classrooms, we aimed to establish a clearer picture of benefits from retrieval practice in real world educational settings. We developed specific search syntax and operationalized classroom research, investigated unresolved questions in the research literature, and developed research-based recommendations for the implementation of retrieval practice for educators.

Methods

Literature Search

We conducted a literature search in January 2018 for empirical research on retrieval practice conducted in school and classroom settings. We developed search syntax for five databases (PsycINFO, PsycARTICLES, ERIC, Web of Science, and PubMed), using different combinations of keywords including retrieval practice, testing effect, course, and teach (see Table 1 for a complete description of the search syntax used for each database). We also performed a backwards search using the reference lists of the abstracts that passed both initial and detail screenings, but the backwards search did not reveal any new abstracts. Our literature search yielded a total of 1,810 abstracts. We used Zotero (http://www.zotero.org), an open-source research tool for reference management, to organize and download abstracts.

Table 1

Search syntax used and number of abstracts screened

Database	Search syntax	Number of Abstracts
PsycINFO, PsycARTICLES, and ERIC	(retriev* pract* OR testing effect OR "test* effect" OR test-enhanc*) AND (class* OR course* OR teach* OR clinical)	1,447
Web of Science	(("retrieval practice" OR "testing effect" OR test-enhanc*) AND (class* OR course* OR teach*))	230
PubMed	(((((retrieval practice[Text Word]) OR testing effect[Text Word]) OR test effect[Text Word]) OR test-enhanced[Text Word])) AND ((((class[Text Word]) OR course[Text Word]) OR teach[Text Word]) OR clinical[Text Word])	133
Total number of abstracts screened		1,810

Initial Screening Criteria

Initial screening of 1,810 abstracts was conducted to ensure that research met the following criteria for inclusion in the present review:

Research must be published in, or in press at, a peer-reviewed journal at the time of our search. Abstracts from conference proceedings, dissertations, or non-peer reviewed journals were excluded. Literature reviews were also excluded.

Research must be empirical, with at least two conditions (i.e., a retrieval practice intervention and a comparison) and at least one final test phase (i.e., a retention measure after retrieval practice). Abstracts indicating that findings were based on observational, survey, correlational, or other qualitative methods were excluded.

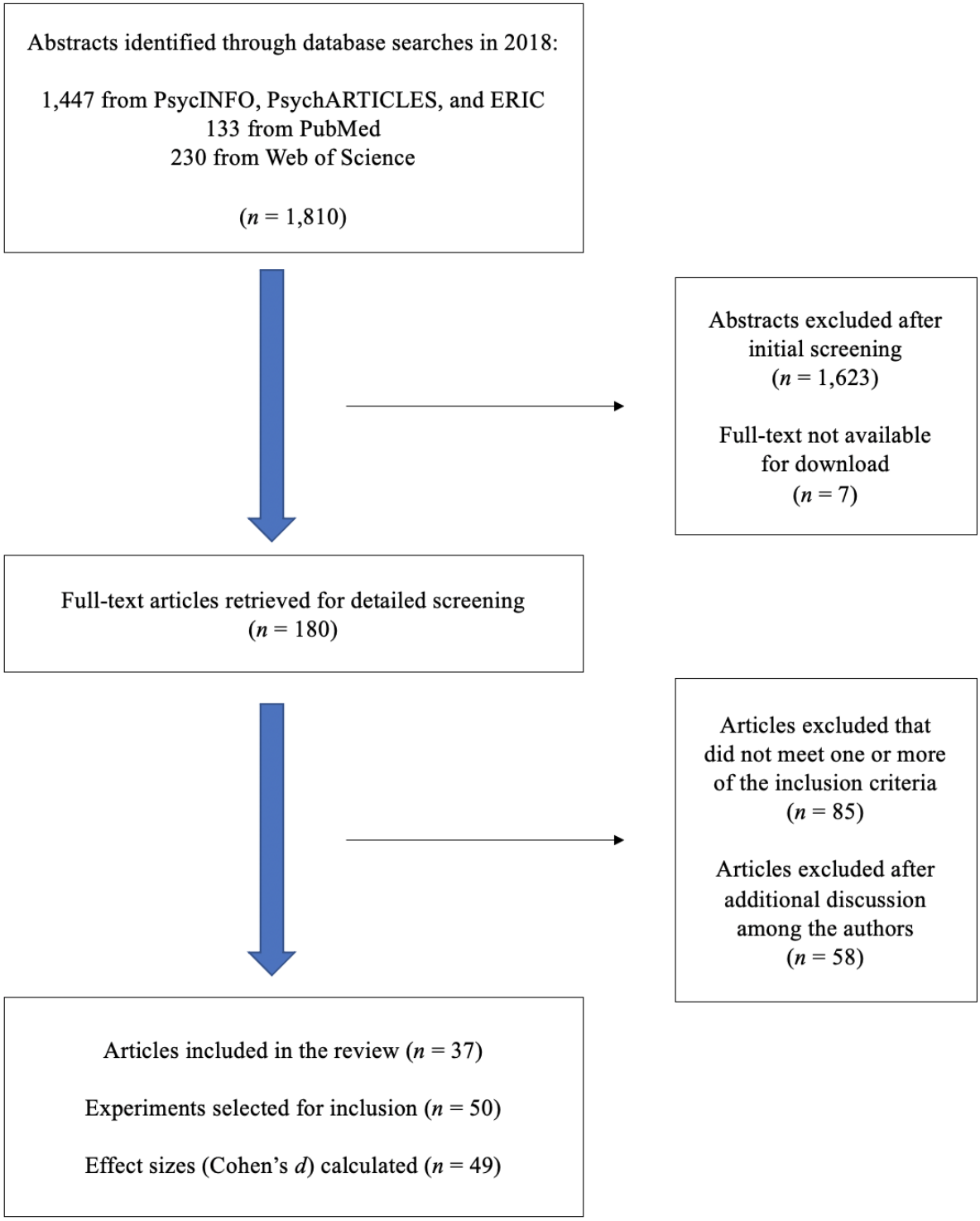
Research must be conducted with typical student populations. Abstracts indicating that research was conducted with patient populations were excluded. For example, participants in Coyne et al. (2015) were traumatic brain injury patients and participants in Viveiros et al. (2017) were patients with heart failure.

Research must include a measurement of retention of information as the dependent variable. Abstracts indicating that the dependent variable was test anxiety or students' preferred study strategies were excluded (e.g., Agarwal et al., 2014; Hartwig & Dunlosky, 2012; Karpicke, Butler, & Roediger, 2009).

As shown in Figure 1, a total of 1,623 abstracts were excluded from the review following initial screening. Seven abstracts that were not available for full-text download were also excluded during initial screening.

Figure 1

Flowchart of the screening process



Detailed Screening

Detailed screening of the remaining 180 abstracts was conducted using the following criteria:

Procedurally, all instructional activities, retrieval practice, and assessments must take place in person during class periods under the supervision of the instructor or researcher. For example, research conducted in laboratories (Lipko-Speed et al., 2014) or online (Becker-Blease & Bostwick, 2016) was excluded. Retrieval practice carried out after the course ended was also excluded (e.g., Carpenter et al., 2009). Furthermore, retrieval practice must be completed by individual students (i.e., research conducted in collaborative groups was excluded; e.g., Vojdanoska et al., 2010) and controlled by the instructor (i.e., research where retrieval practice was self-regulated using flashcards was excluded; e.g., Rawson, Dunlosky, & Sciartelli, 2013).

In terms of materials, information to be learned must be the same as, or directly related to, assigned course materials that students would be learning in the absence of researchers. For example, Duchastel (1979) had students memorize passages about solar power that were not part of course materials; thus, it was excluded from the present review. In addition, the retrieval practice intervention must take place without the use of notes, external learning aids, or the internet (e.g., quizzes in class must be closed-book, not open-book). In addition, the amount of instructional time during which students were exposed to materials must be equivalent across retrieval practice and comparison conditions.

Coding Procedure

Following initial and detailed screenings, 50 experiments drawn from 37 studies were included in the present review. As shown in the Appendix, we coded the following variables for each of the 50 experiments that passed all screening criteria: (a) the type of retrieval practice

intervention; (b) the comparison conditions; (c) calculated effect sizes and confidence intervals; (d) education level (i.e., K-12, undergraduate, or medical school); (e) content area (e.g., science, psychology, history, etc.); (f) specific course topic (e.g., biology); (g) the experimental design; (h) sample size after attrition; (i) whether the experiment was conducted in the United States; (j) retrieval practice timing; (k) the delay between the last retrieval practice opportunity and the final test; (l) the format of retrieval practice (e.g., multiple-choice or short answer); (m) the provision of feedback after retrieval practice (e.g., immediate or delayed); (n) the format of the final test; and (o) whether final test performance counted toward students' grades. When any variables to be coded were ambiguous, two or more of the present authors coded the experiment independently and resolved discrepancies.

Effect Size Calculations

When coding or calculating effect sizes, we used performance on the final test that occurred in closest proximity to the last instance of retrieval practice to avoid practice effects. For example, in Roediger et al. (2011), middle school students completed retention tests at the end of each chapter and also at the end of the semester; thus effect sizes from Roediger et al. are reported based on chapter test performance only.

Across the 50 experiments coded, we derived 49 effect sizes (Cohen's *d*). For 24 comparisons, data were insufficiently reported to calculate an effect size. Whenever possible, we calculated the effect sizes and corresponding confidence intervals around the effect sizes. We did so even when the original study reported effect sizes because (a) only one study reported confidence intervals around effect sizes (McDermott et al., 2014) and (b) this ensured consistency in the way the effect sizes were calculated. Note that our calculated effect sizes did not match all of the effect sizes reported in the original articles, although that appears to be

common as researchers use different formulas to calculate effect sizes (Pan & Rickard, 2018). This is true especially for within-subject designs, for which the correlation between data must be accounted for.

For between-subjects designs, effect sizes and respective 95% confidence intervals were calculated from a reported or derivable *t* statistic and a reported or derivable sample size, or from reported or derivable sample sizes, means, and standard deviations. For within-subject designs, effect sizes were calculated from a reported or derivable *t* statistic and a reported or derivable sample size. To compute the effect sizes and 95% confidence intervals around them, we used the MBESS package for R (functions ci.sm and ci.smd for within- and between-subject designs, respectively; Kelley, 2007a, 2007b, 2017).

We focus our discussion of effect sizes on our calculated Cohen's *d*s and we categorized obtained effect sizes as large, medium, and small using Cohen's (1988) standards. Large effect sizes were defined as $d > 0.80$; medium effects were $0.50 < d < 0.80$; small effects were $0.20 < d < 0.50$; very small effects were $0.00 < d < 0.20$; and negative effects were $d < 0.00$. The forest plots depicted in Figures 3-6 allow for a visual representation of the effect sizes and their precision (i.e., narrower confidence intervals indicate more precise effects than wider confidence intervals).

Results

Coding for all experiments is available in the Appendix, as well as on the Open Science Framework (http://osf.io/mz2ks/). Across the 50 experiments coded in the present review, the total sample size was $n = 5{,}374$ (sample size not reported for Graham, 1999). Altogether, experiments coded ranged across a number of factors:

- Education levels ranged from elementary school to medical school (e.g., Goossens et al.,

2016 and Larsen et al., 2013a, respectively)

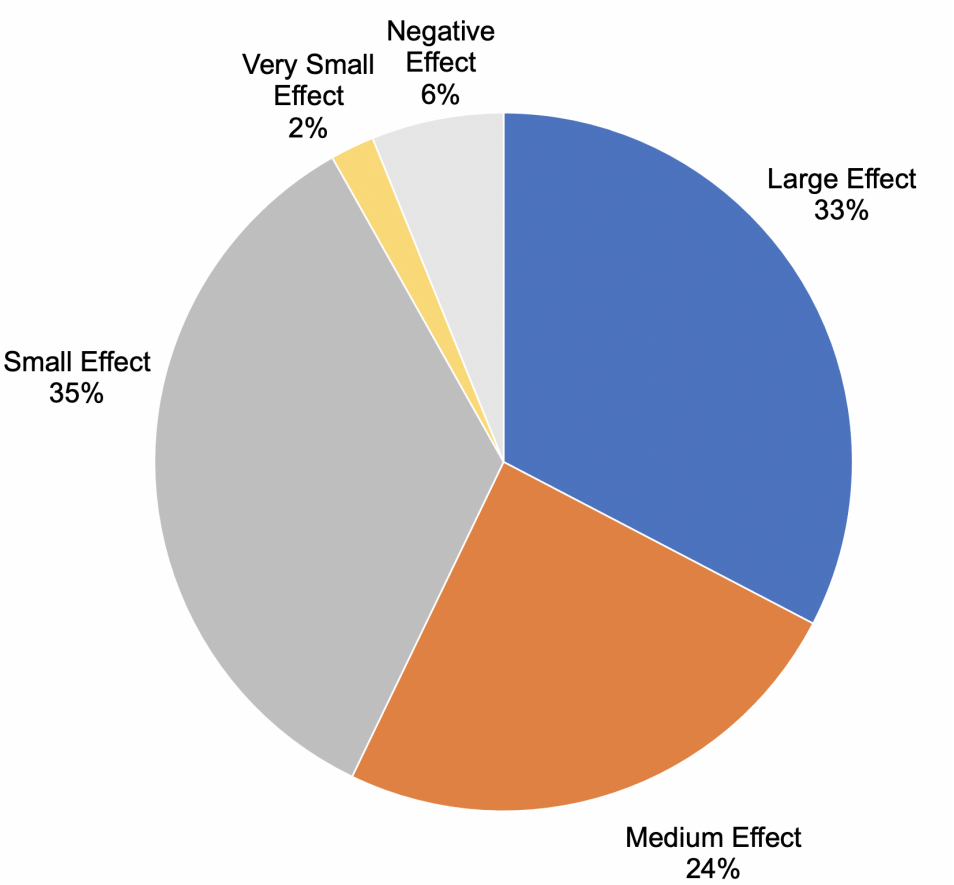
- Sample populations included three non-WEIRD countries: Pakistan (e.g., Ayyub & Mahboob, 2017), Taiwan (e.g., Tu et al., 2017), and Turkey (e.g., Atabek Yigit et al., 2014)
- Sample sizes after attrition ranged from fewer than 20 students to nearly 400 students (Jones et al., 2016 and Bjork et al., 2014, respectively)
- Delays between retrieval practice and the final test ranged from one day to the end of the semester (e.g., McDaniel et al., 2013 and Tu et al., 2017, respectively)

As displayed in Figure 2, the majority of effect sizes (57%) indicated medium or large benefits from retrieval practice. In other words, 28 out of 49 Cohen's *d*s were greater than 0.50. Overall, 16 effect sizes indicated large benefits from retrieval practice ($d > 0.80$), 12 indicated medium benefits ($0.50 < d < 0.80$), and 18 were small or very small ($d < 0.50$). Only three out of 49 effect sizes revealed a negative effect, or a benefit for the comparison condition (lessons without quizzes) compared to retrieval practice (Khanna, 2015; Michaels, 2017; Tu et al., 2017).

In Figure 3, all 49 effect sizes are depicted in a forest plot, which includes the 95% confidence intervals around each effect size. The effect sizes are organized from the largest to the smallest, and the width of the confidence intervals represents the precision of the effect size estimate. The confidence intervals vary widely, possibly because of variability in sample size across studies. However, only six confidence intervals around positive effect sizes extend below a Cohen's *d* of 0.00. This suggests that for almost all studies reviewed, possible values for effect sizes are in a positive direction, indicating a consistent benefit from retrieval practice on student learning.

Figure 2

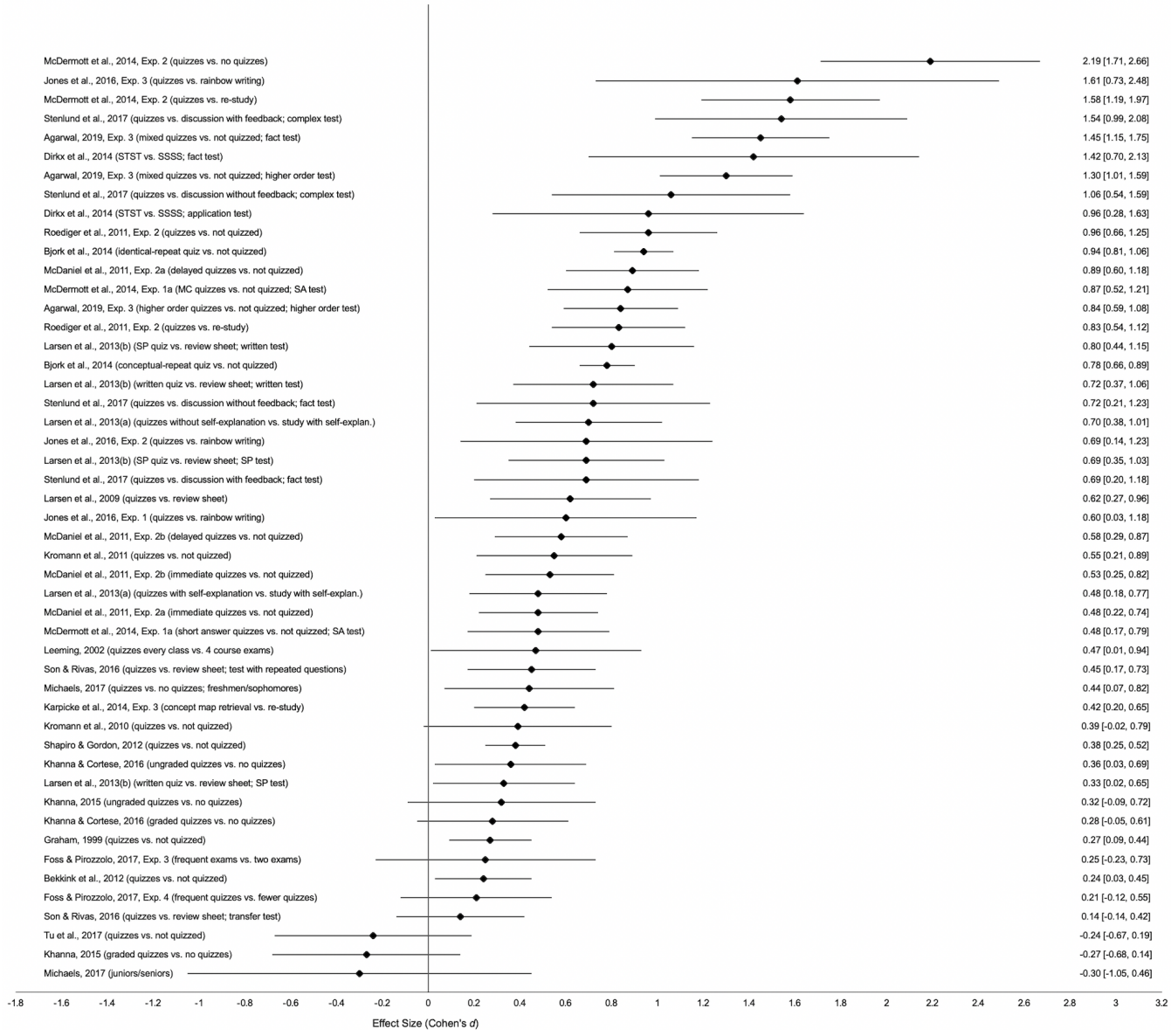
Distribution of 49 effect sizes (Cohen's d) from the articles reviewed



Note. Large effect sizes were defined as $d > 0.80$; medium effects were $0.50 < d < 0.80$; small effects were $0.20 < d < 0.50$; very small effects were $0.00 < d < 0.20$; and negative effects were $d < 0.00$ (Cohen, 1988). See the Appendix for a complete list of effect sizes for each experiment.

Figure 3

Forest plot of 49 effect sizes (Cohen's d) from the articles reviewed



Moderating Variables: Education Level, Content Area, Comparison Conditions, Experimental Design, Sample Size, and Location

As shown in Table 2 and the Appendix, experiments were evenly distributed across education levels. Of the 50 experiments reviewed, 20 experiments were conducted in K-12 settings (elementary, middle, and high school); 20 experiments were conducted in undergraduate settings (college/university level); and 10 experiments were conducted in medical schools.

As shown in Figure 4, effect sizes were largest for studies conducted in middle school classrooms. Effect sizes ranged from small to large for medical school, and studies conducted at the undergraduate level resulted in smaller effect sizes. Note that all of the studies at the middle school level were conducted within the same school district near St. Louis, Missouri, United States (Agarwal, 2019; McDaniel et al., 2011; McDaniel et al., 2013; McDermott et al., 2014; Roediger et al., 2011). Of the eight studies at the medical school level, three were conducted at a medical school in Copenhagen, Denmark (Kromann et al., 2009, 2010, 2011) and three were conducted at a medical school in St. Louis, Missouri (Larsen et al., 2009, 2013a, 2013b). The nine studies at the undergraduate level were mostly conducted at different colleges and universities.

Regarding content area, most experiments were conducted in science (k = 19) and psychology courses (k = 16), with few experiments conducted in history (k = 5), skills-based learning (k = 5), spelling and vocabulary (k = 4), and statistics (k = 2; one experiment included both science and history). These results indicate that more applied research is needed in non-science areas, particularly in skills-based learning, mathematics, the humanities, and foreign language learning.

In contrast to retrieval practice, the most common comparison conditions were when

students re-read material (e.g., studied a review sheet; $k = 19$), when the instructor provided lessons without retrieval practice ($k = 14$), and when lessons included retrieval practice, but comparison performance was measured on non-quizzed items on a final test ($k = 12$). Only five experiments included a comparison condition of fewer opportunities for retrieval practice (e.g., comparing two exams vs. weekly quizzes); experiments using this type of comparison yielded small and very small effect sizes. Considering recent recommendations for educators to provide frequent retrieval practice in their already established lessons and course structure (Agarwal & Bain, 2019), more research needs to examine the extent to which the quantity of retrieval practice modulates benefits on learning compared to infrequent exams.

In sum, the majority of experiments revealed medium to large effect sizes, indicating that retrieval practice consistently improves learning in schools and classrooms for a variety of education levels and content areas, under diverse comparison conditions.

As shown in Table 2 and the Appendix, a majority of experiments were conducted within-student ($k = 29$). Twelve experiments were conducted between-students without random assignment, and nine experiments were conducted between-students with random assignment. Considering the logistical challenges and ethical concerns of random assignment in applied school settings, we were surprised to find that nearly half of the between-subjects experiments included random assignment.

Within-student experiments revealed a range of effect sizes, with more medium to large effect sizes than small effect sizes. Between-student experiments with random assignment tended to have larger effect sizes, while experiments without random assignment had smaller effect sizes. These results indicate that, similar to laboratory studies, retrieval practice improves learning in applied settings whether experiments are conducted within-student or between-

students, with or without random assignment.

Regarding sample size, the majority of experiments (64%) were conducted with fewer than 100 students (see Table 2 and the Appendix). Average sample size varied by education level: K-12 (M = 57.7), undergraduate (M = 149.3), and medical school (M = 134.8). Average sample size also varied by experimental design: within-student (M = 78.2), between-students (M = 153.3), and between-students with random assignment (M = 149.6). As shown in Figure 4, experiments conducted at the undergraduate level tended to have the smallest effect sizes, but these studies also had the largest sample sizes. It is possible that retrieval practice may simply be more beneficial for middle school and medical school students, compared to undergraduate students, which we consider further in the General Discussion.

The vast majority of experiments reviewed (94%) were conducted in the United States and Western Europe, consistent with prior findings that the majority of published studies in psychology are conducted in WEIRD countries (western, educated, industrialized, rich, democratic countries; Rad, Martingano, & Ginges, 2018). Only three out of 50 experiments (6%) drew samples from schools outside the United States and Western Europe, representing 6% of the total sample size in our review (n = 266 out of n = 5,374). The three experiments conducted in non-WEIRD countries were from Pakistan, Taiwan, and Turkey. Of these three, one experiment was conducted at the undergraduate level and two were conducted at medical schools. In other words, all experiments at the elementary, middle, and high school levels included in our review were conducted in WEIRD countries. To foreshadow our General Discussion, applied research on retrieval practice is needed with diverse student populations from non-WEIRD countries in order to provide accurate recommendations for educators globally.

Additional research with larger and more diverse sample sizes would further our understanding of the benefits of retrieval practice in educational settings. In addition, measures of individual differences could be included to examine whether there are optimal conditions for retrieval practice depending on, for example, working memory, prior knowledge, intelligence, and mind wandering (Agarwal, Finley, Rose, & Roediger, 2017; Francis, Wieth, Zabel, & Carr, 2020; Minear, Coane, Boland, Cooney, and Albat, 2018; Pachai, Acai, LoGiudice, & Kim, 2016).

Table 2

*Distribution of experiments (*k*) by moderating variables*

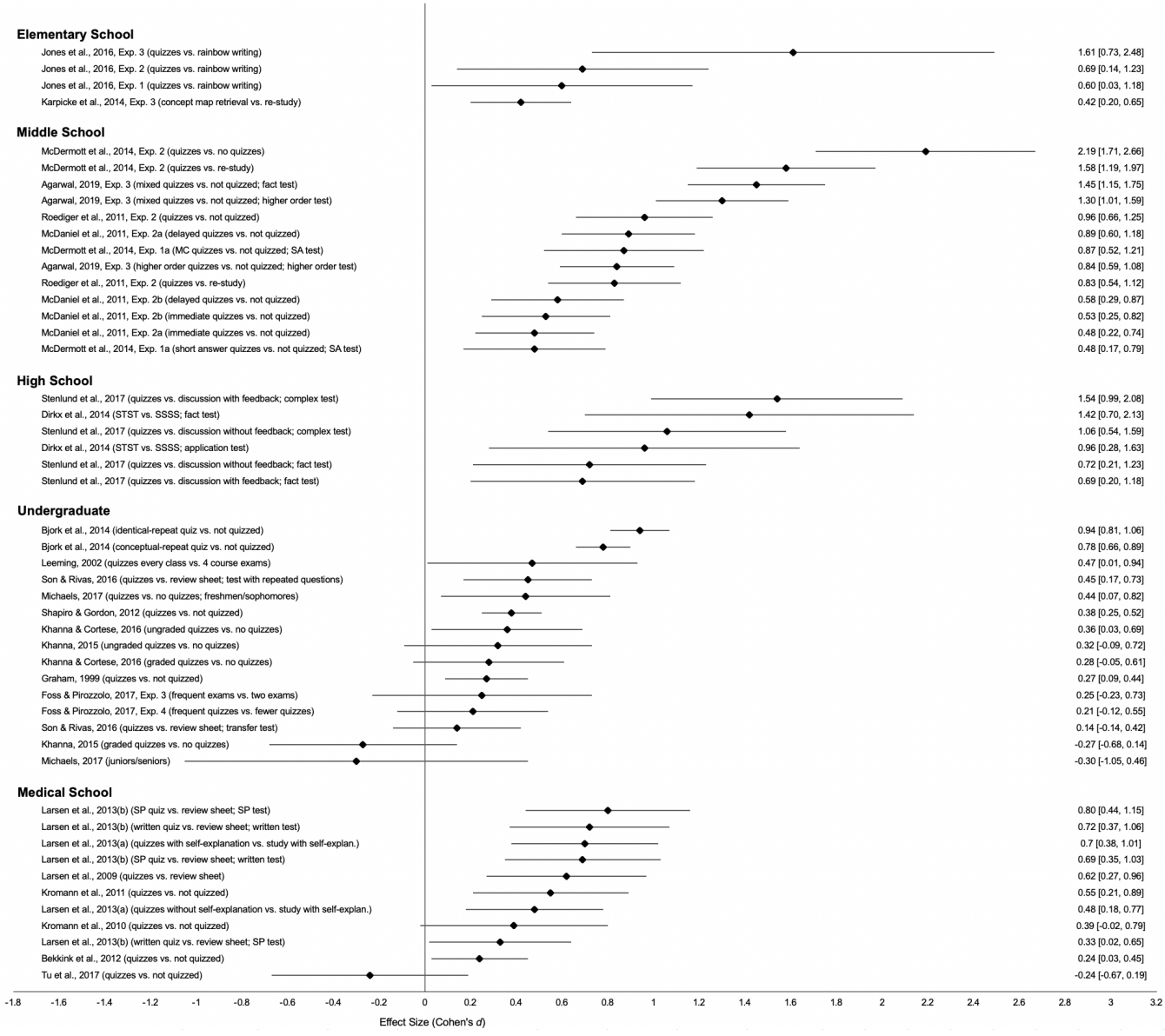
	Total number of experiments
Education level (k = 50)	
Elementary school	5
Middle school	12
High school	3
Undergraduate	20
Medical school	10
Content area (k = 51)	
Science	19
Psychology	16
History	5
Skills-based (CPR, dental diagnosis, nursing skills)	5
Spelling and vocabulary	4
Mathematics (statistics)	2
Type of comparison condition (k = 50)	
Re-read material	19
Lessons without quizzes	14
Non-quizzed items on the final test	12

Fewer opportunities for retrieval practice	5
Experimental design (k = 50)	
Within-student	29
Between-students without random assignment	12
Between-students with random assignment	9
Sample size (k = 49)	
Fewer than 100 students	32
Greater than 100 students	17
Location (k = 50)	
WEIRD countries (United States, Denmark, Netherlands, Sweden)	47
Non-WEIRD countries (Pakistan, Taiwan, Turkey)	3

Note. The total number of experiments reported under content area is k = 51 because Karpicke et al. (2014, Experiment 3) included content from both science and history. The total number of experiments reported under sample size is k = 49 because it was not reported by Graham (1999). When an experiment included more than one comparison condition, the number of experiments listed refers to the type of condition in which time spent with material was similar or equivalent to time spent engaged in retrieval practice.

Figure 4

Forest plot of effect sizes (Cohen's d) by education level



Retrieval Practice Timing and Delay Before the Final Test

As shown in Table 4, retrieval practice was typically provided at least once per week (k = 19) or every 2-3 weeks (k = 15). In a few experiments (k = 6), retrieval practice was provided multiple times throughout the semester, but specific timing was not reported. In the remaining experiments, retrieval practice was provided within a single session (k = 10). Because effect sizes were evenly distributed across a range of timings (see Appendix), we recommend educators provide students with opportunities for retrieval practice regardless of the precise timing.

Table 4

*Distribution of experiments (*k*) by retrieval practice timing and final test delay*

	Total number of experiments
Retrieval practice timing (k = 50)	
Single session	10
At least once per week	19
At least once every 2-3 weeks	15
Multiple times throughout the semester (duration not specified)	6
Delay between retrieval practice and final test (k = 50)	
Immediate	4
1-3 day delay	20
1-2 week delay	7
Multiple exams throughout the semester (timing not specified)	5
One exam at the end of the semester or course (6-15 week delay)	14

The most common delay between the last opportunity for retrieval practice and the final test was a 1-3 day delay (k = 20; see Table 4). The next most frequent delay was when a final exam occurred at the end of the semester or conclusion of the course, after approximately 6-15 weeks, although a specific delay was not reported (k = 14). Effect sizes were larger following a 1-3 day delay, while smaller following an end-of-semester delay (see Appendix). In other words, shorter delays led to a larger benefit from retrieval practice in classroom settings. However, in laboratory research, the opposite effect has been shown—*longer* delays lead to a larger benefit (Carpenter & Agarwal, 2020; Roediger & Karpicke, 2006a). For both theoretical and practical considerations, we encourage future research where a range of delays between retrieval practice and final tests are directly manipulated.

Retrieval Practice Format and Final Test Format

As shown in Table 5, the most common formats for retrieval practice were multiple-choice (k = 27) and short answer (k = 17). Similarly, the majority of final test formats were multiple-choice (k = 31) and short answer (k = 15). As displayed in Figure 5, effect sizes were larger when retrieval practice and final test formats matched (multiple-choice or short answer). When an experiment included multiple formats (e.g., multiple-choice retrieval practice followed by a short answer final test; k = 11), effect sizes were smaller. The transfer appropriate processing framework may account for these findings, where a match between initial and final processing typically promotes learning (Morris, Bransford, & Franks, 1977). As such, we recommend both multiple-choice and short answer formats for retrieval practice, and a match with final test format may be optimal for promoting student learning.

Table 5

*Distribution of experiments (*k*) by retrieval practice format and final test format*

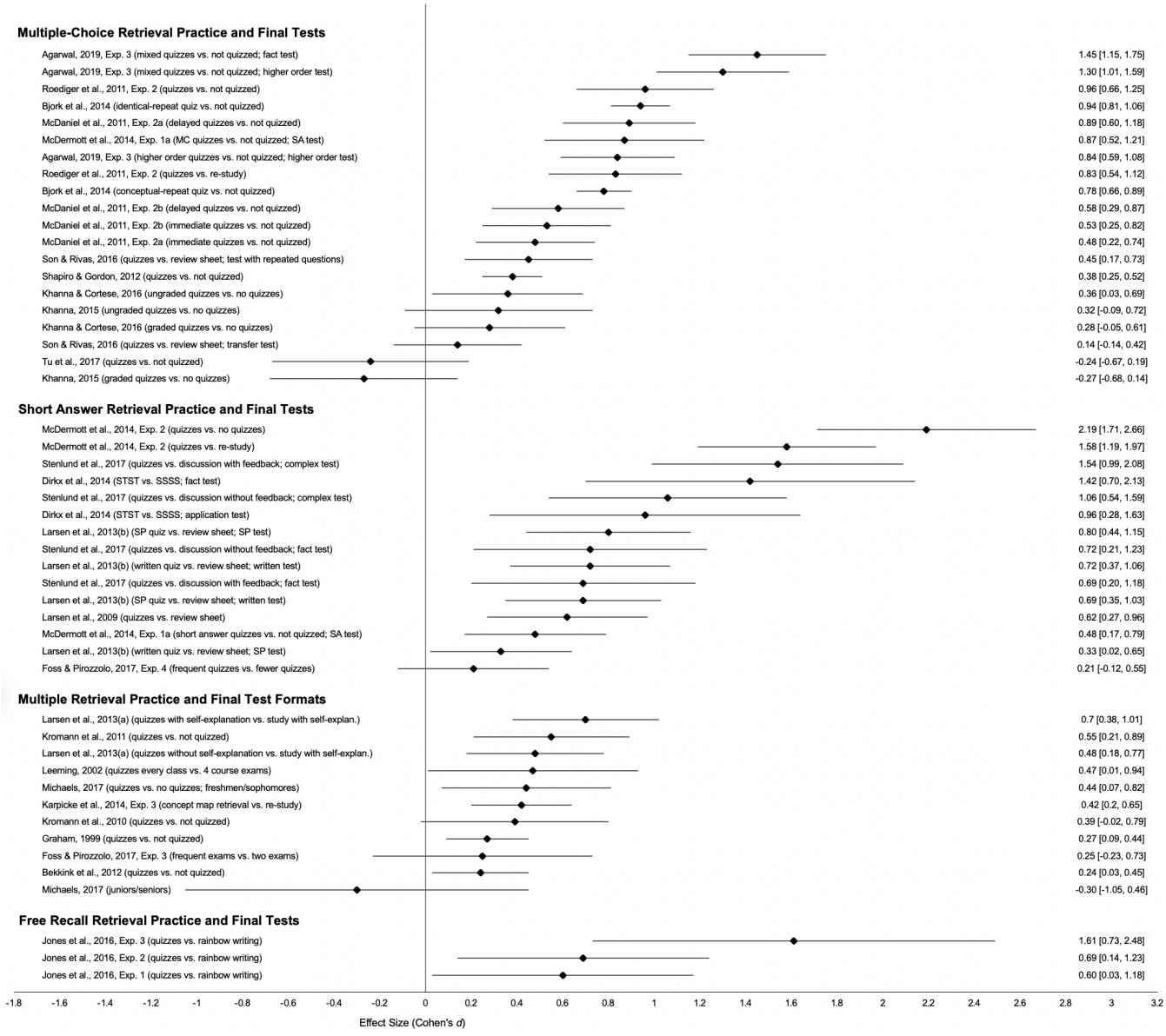
	Total number of experiments
Retrieval practice format (k = 62)	
Multiple-choice	27
Short answer	17
Free recall	6
Cued recall, fill-in-the-blank, or matching	6
Simulated diagnoses	5
Retrieval format not reported	1
Final test format (k = 64)	
Multiple-choice	31
Short answer	15
Free recall or essay	9
Cued recall, fill-in-the-blank, or matching	4
Simulated diagnoses	5
Final test questions (k = 50)	
Rephrased	22
Verbatim	28

Final test scores counted toward students' grades (k = 50)	
Yes	33
No	15
Not reported or unavailable	2

Note. The total number of experiments reported for retrieval practice format and final test format is greater than k = 50 because some experiments included multiple formats.

Figure 5

Forest plot of effect sizes (Cohen's d) by retrieval practice and final test formats



When students engage in retrieval practice, a common concern is that they are simply learning the test questions and answers (i.e., a practice effect). Thus, we investigated whether questions during retrieval practice were typically rephrased or verbatim on the final test. For nearly half of the experiments reviewed, questions were rephrased (k = 22), which was most common at the undergraduate and medical school levels. Effect sizes were generally smaller for experiments with rephrased questions compared to experiments with verbatim or repeated questions. Because transfer of knowledge following retrieval practice remains a challenge in the classroom and in the literature (Agarwal, 2019; Butler, 2010; Pan & Rickard, 2018), we encourage more research on retrieval practice and transfer specifically in applied settings. Transfer is, after all, a "holy grail" of education (Pan & Agarwal, 2020).

The majority of experiments included final test performance as part of students' grades (k = 33; Table 5). Effect sizes ranged from small to large regardless of whether test scores were included as part of students' grades, indicating that retrieval practice improved student learning under typical motivational factors in classroom settings, supporting prior laboratory research (Abel & Bäuml, 2020; Kang & Pashler, 2014).

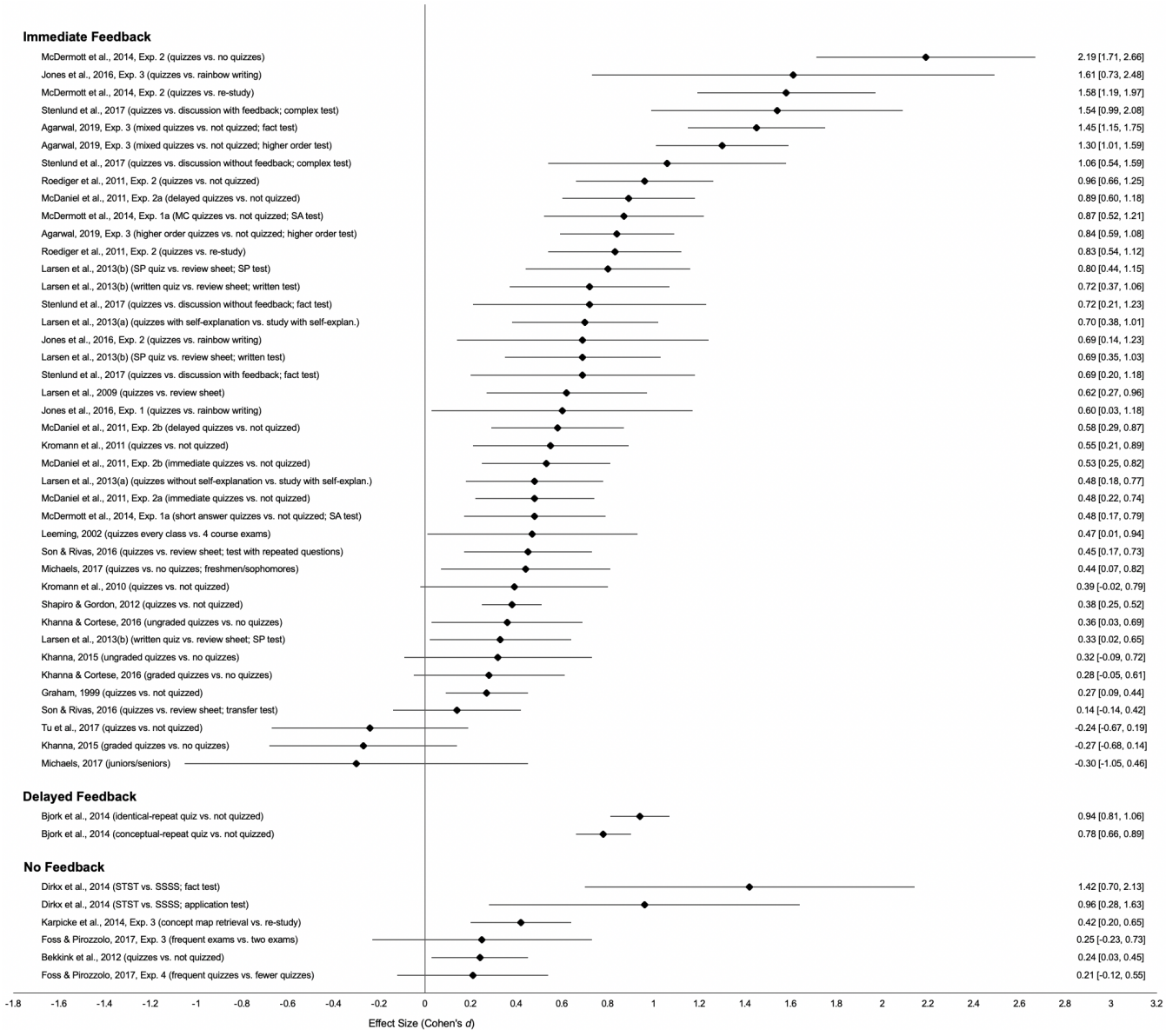
Feedback Provision and Timing

The majority of experiments provided immediate feedback (k = 34). Ten experiments did not include feedback and only four experiments included delayed feedback; two studies did not report whether feedback was provided. In Figure 6, effect sizes for immediate feedback are evenly distributed across small, medium, and large effects. Four studies included delayed feedback, but only one study had data available to calculate Cohen's ds (Bjork et al., 2014), which showed medium to large effects. Studies without feedback resulted in mostly small or very small effects on learning.

Due to limited reporting in the studies reviewed, we were unable to code for whether feedback was administered after each initial question, at the end of the quiz, or at the end of the class session. We were also unable to code for the type of feedback (e.g., correct answer, elaborative, etc.) due to lack of reporting. Thus, we were unable to establish an optimal timing of feedback in school and classroom settings.

Figure 6

Forest plot of effect sizes (Cohen's d) by timing of feedback provided



General Discussion

Does retrieval practice improve student learning in school and classroom settings? Based on our literature review, our response for researchers and educators is an unequivocal "yes." We found a wealth of evidence, based on medium to large effect sizes, that retrieval practice improved learning for a variety of education levels, content areas, experimental designs, retrieval practice timing, final test delays, retrieval and final test formats, and the timing of feedback.

Compare Apples to Apples Using Narrower Review Criteria

Our first aim for the present literature review was to compare "apples to apples." In other words, we focused our review on retrieval practice research conducted specifically in school and classroom settings, and we intentionally omitted research conducted in laboratory settings. After screening nearly 2,000 abstracts, we narrowed our literature review to 37 studies with 50 experiments and 49 total effect sizes, with a total n = 5,374. The majority of effect sizes (57%) from studies reviewed were medium or large (Cohen's d; Figure 2).

As is the case for all systematic reviews, ours is subject to publication bias, when studies with positive outcomes tend to be published to a greater extent than studies with negative outcomes (Augusteijn, van Aert, & van Assen, 2019; Simmons, Nelson, & Simonsohn, 2011). As stated by Ferguson and Heene (2012), "[publication bias] is a systemic discipline-wide problem" in psychology. Specific to research on retrieval practice in classrooms, our search did not return any studies published before 1999 that met our criteria for inclusion. Furthermore, we did not include unpublished studies, as the inclusion of unpublished studies can be ill-defined and thus not reduce publication bias (Ferguson & Heene, 2012). Also, new research on retrieval practice in classroom settings has been published since our search in January 2018 (e.g., Gurung et al., 2019). We did, however, find a number of studies that were not included in previous

reviews of research on retrieval practice. We hope that by making our literature review and Appendix publicly available (http://osf.io/mz2ks), we will increase access to research in this field for both researchers and educators, a small step toward addressing publication bias.

Future Directions for Research on Retrieval Practice

Our second aim was to inform future research examining retrieval practice. Following our review of the literature, we provide eight recommendations.

First, the field needs more applied research investigating varying delays between retrieval practice and the final test. We found larger effect sizes at shorter delays (1–3 days) and smaller effect sizes at longer delays (end of the semester). In other words, results from our review of the literature indicate the opposite of what is typically found in laboratory studies, where benefits from retrieval practice are *larger* after longer delays (Carpenter & Agarwal, 2020; Roediger & Karpicke, 2006a). One possible explanation for this discrepancy is that the delays in classroom settings are much longer than in laboratory settings. In the review of the literature on retrieval practice by Adesope and colleagues (2017), the majority of studies (64%) had a delay of six or fewer days. Therefore, it may be the case that benefits from retrieval practice increase in the first few days, and then taper off as delays approach weeks or months. For both theoretical and practical considerations, we encourage future research where a range of delays between retrieval practice and final tests are directly manipulated, ideally ranging from days to weeks to months. Courses administered online may be particularly suitable for this type of research.

Second, classroom research specifically investigating the provision and timing of feedback is needed. Although feedback is a key component of educational settings, classroom studies that directly manipulated feedback (e.g., immediate vs. delayed) were notably absent from our literature search. The benefits of immediate vs. delayed feedback also remain unclear in

laboratory research (Agarwal, Bain, & Chamberlain, 2012; Kulik & Kulik, 1988; Metcalfe, Kornell, & Finn, 2009; Mullet, Butler, Verdin, von Borries, & Marsh, 2014). As was the case for our first recommendation, it is of both theoretical and practical importance to fully understand whether immediate or delayed feedback produces the largest benefits for student learning. One consideration to keep in mind is that delayed feedback can present logistical challenges in classroom settings. Online classes, on the other hand, provide an opportunity to examine delayed feedback with fewer logistical challenges (Butler, Marsh, Slavinsky, & Baraniuk, 2014).

Third, comparison conditions for future research on retrieval practice should more closely mirror common classroom practices. We found that many studies in our literature review compared retrieval practice to re-reading. It has been well-established in both laboratory and applied research (including research in the current review) that retrieval practice significantly increases student learning when compared to re-read exposure controls (e.g., Atabek Yigit et al., 2014; Dunlosky et al., 2013; McDermott et al., 2014; Roediger & Karpicke, 2006a). More stringent and realistic comparisons to retrieval practice in classroom settings include traditional lectures, flipped classroom activities, think-pair-share discussions, and student presentations (DeLozier & Rhodes, 2017).

Fourth, future research should also examine factors unique to applied settings including class size (e.g., lecture classes vs. small classes), whether retrieval practice was anticipated (e.g., "pop quiz" vs. announced in advance), whether the final test was cumulative, and whether performance on initial retrieval practice counted towards students' grades. We coded whether final test performance counted toward students' grades, but information regarding grading procedures for initial retrieval practice was not reported in most studies.

Fifth, additional research is needed in non-science content areas, such as skills-based

learning, mathematics, the humanities (writing, literature, essays), and foreign language vocabulary. Thirty-five out of the 50 experiments reviewed were conducted in science or psychology courses. We were particularly surprised that none of the experiments meeting our screening criteria included foreign language learning, considering the frequent use of these materials in laboratory experiments (Dunlosky et al., 2013).

Sixth, applied research in education should also take into account the role of the teacher-researcher as a modulating factor for student learning outcomes, also known as the Hawthorne Effect or "participant reactivity" (Diaper, 1990; Paradis & Sutkin, 2017). This information was inconsistently reported in the studies reviewed, unfortunately. When reported, we found that instructors at the undergraduate level tended to also be the researchers (e.g., Batsell et al., 2017; Leeming, 2002; Lyle & Crawford, 2011; Saville et al., 2012), whereas instructors at the K-12 level were not the researchers (e.g., Agarwal, 2019; Karpicke et al., 2014; McDaniel et al., 2011, 2013; McDermott et al., 2014; Roediger et al., 2011). As such, the role of the teacher-researcher may have contributed to smaller effect sizes in undergraduate classrooms (Figure 4). Another possibility for smaller effect sizes at the undergraduate level may be due to larger sample sizes or the more frequent use of rephrased questions on final tests. Follow up studies specifically comparing K-12, undergraduate, and medical school students could shed light on whether benefits from retrieval practice are modulated by age in applied settings.

Seventh, collaborative retrieval and online quizzes are common in educational settings and it would be beneficial to know when and how they increase student learning. While conducting our literature search, we found numerous educational studies conducted under online or collaborative conditions. While these studies were outside the scope of our review (both areas of research warrant their own literature reviews), they can add to our understanding of retrieval

practice in real world settings. With online learning, for example, instructors have more control and flexibility over the provision, timing, and frequency of retrieval practice and feedback (Butler, Marsh, Slavinsky, & Baraniuk, 2014). With collaborative retrieval, a literature review could highlight optimal collaborative groups for a range of ages, content areas, and metacognitive skills (de Carvalho Filho, 2010).

As our final recommendation, applied research on retrieval practice must be conducted with diverse student populations. We found that only three out of 50 experiments that met our screening criteria were conducted outside the United States and Western Europe (Turkey, Pakistan, and Taiwan), while 94% of classroom research on retrieval practice was conducted in WEIRD countries (western, educated, industrialized, rich, democratic countries; Henrich, Heine, & Norenzayan, 2010). The discrepancy is clear: non-WEIRD countries account for 88% of the global population, but only 6% of the sample from the studies reviewed were from non-WEIRD countries (Bauer, 2019; Rad, Martingano, & Ginges, 2018).

Based on findings from the present review—retrieval practice consistently improved learning across a range of ages, content areas, formats, etc. in WEIRD countries—it stands to reason that retrieval practice would similarly benefit student learning in non-WEIRD countries. In line with this reasoning, recent research conducted in non-WEIRD countries suggests that retrieval practice improves learning for elementary school children in Brazil (de Lima & Jaeger, 2020) and also for college students in Hungary (Racsmány, Szőllősi, & Bencze, 2018), countries that are considered to be low in educational attainment (OECD, 2020). In addition, self-reported study strategies used by students in Brazil are similar to study strategies used by students in the United States (Ekuni, de Souza, Agarwal, & Pompeia, 2020), and researchers have found limited cross-cultural differences on measures of working memory (Adams & Hitch, 1997; Lan, Legare,

Ponitz, Li, & Morrison, 2011), a cognitive process engaged during retrieval practice (Agarwal et al., 2017).

While research suggests that retrieval practice may benefit long-term learning for all learners (WEIRD and non-WEIRD), *implementation* by students and educators may be affected by cultural norms. For example, van Egmond, Kühnen, and Li (2013) found that the definition of academic learning varies by culture: in Western cultures, learning is attributed to the cognitive domain, whereas in Eastern cultures, learning is associated with the development of the person as a whole; thus, it may be the case that retrieval practice could be implemented more frequently in Western cultures consistent with a cognitive approach. Tweed and Lehman (2002) suggested that ideals of learning that are predominantly Western (Socratic) or more Eastern (Confucian) influence students' approaches toward learning, including motivation, effort, and memorization, all of which are contributing factors to the implementation of retrieval practice (Agarwal & Bain, 2019). Furthermore, Western cultures are considered individualistic or independent because they focus on standing out and being unique, whereas Eastern cultures are considered collectivist or interdependent because they focus on maintaining harmony within the group (Markus & Kitayama, 1991); to speculate, this could affect the extent to which educators implement individual vs. collaborative retrieval practice. Considering these cultural dynamics, it is possible that the implementation of retrieval practice in WEIRD countries may be vastly different from implementation in non-WEIRD countries, subsequently modulating benefits from retrieval practice on student learning.

Lastly, we encourage research on retrieval practice with non-WEIRD students because an overreliance on WEIRD samples can produce false claims about human psychology and behavior (Henrich, Heine, & Norenzayan, 2010). For example, Henrich and colleagues found

that individuals from WEIRD countries exhibit divergent behaviors compared to the rest of the world, even for domains that were previously considered to be universal, such as visual perception, cooperation, spatial reasoning, and moral reasoning. As a second example, the autobiographical reminiscence bump is considered to be a basic memory process, and yet the content of memories differs for individualistic vs. collectivist cultures (Conway, Wang, Hanyu, & Haque, 2005). Third, Segall, Campbell, and Herskovits (1966) found that the San foragers of the Kalahari were immune to the well-known Muller-Lyer visual illusion, which was previously thought to be fundamental to the human species. If these researchers had not tested non-WEIRD samples, we might still hold the conviction that all humans are susceptible to the same patterns of behavior.

For these reasons, it would be shortsighted to assume that what is beneficial for learning in WEIRD countries is beneficial for learning in non-WEIRD countries. Applied research on retrieval practice is needed with students from non-WEIRD countries if we are to provide accurate recommendations, based on empirical evidence, for educators and students globally. As a starting point, we urge researchers to report key demographics, including students' age, gender, location, and type of school (e.g., public, private, rural, urban, etc.). If researchers are to provide practical recommendations for educators regarding retrieval practice, then student demographics must be taken into account (Rad, Martingano, & Ginges, 2018). Educators are eager to know whether retrieval practice would be beneficial for their specific student population, and also whether these benefits generalize to all classrooms.

Recommendations for Classroom Implementation of Retrieval Practice

Our third and final aim for the literature review was to identify practical recommendations for educators as they implement retrieval practice in their classrooms. We had

anticipated the emergence of optimal conditions for retrieval practice: content areas, formats, timing, and so on. However, we did not find any singular or specifically optimal conditions; instead, we found that nearly all conditions in schools and classrooms yielded a benefit from retrieval practice.

We conclude that educators should implement retrieval practice, with less concern about the precise format or timing of retrieval interventions. Almost all effect sizes (46 out of 49 Cohen's *d*) indicated a positive benefit from retrieval practice under wide-ranging conditions, and retrieval practice improved student learning to a greater extent than time spent on other classroom activities (e.g., reviewing material, lectures without quizzes, etc.).

We hope that this literature review provides educators with an accessible resource when considering implementation of retrieval practice. In our Appendix (available at http://osf.io/mz2ks), we have included details for each of the 50 experiments, broken down by education level, content area, effect sizes, and more. Educators can explore the retrieval practice research that has been conducted under similar conditions as their own classroom or school, to inform teaching strategies, professional development, and curriculum development. By implementing retrieval practice in schools and classrooms, scientists and educators can bridge the gap between research and practice—and most importantly, transform students' long-term learning.

References

**References with an asterisk indicate studies included in the review.*

Abel, M., & Bäuml, K. T. (2020). Would you like to learn more? Retrieval practice plus feedback can increase motivation to keep on studying. *Cognition*, *201*. doi.org/10.1016/j.cognition.2020.104316

Adams, J. W., & Hitch, G. J. (1997). Working memory and children's mental addition. *Journal of Experimental Child Psychology*, *67*, 21-38.

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.

*Agarwal, P. K. (2019). Retrieval practice and Bloom's taxonomy: Do students need fact knowledge before higher order learning? *Journal of Educational Psychology*, *111*, 189-209.

Agarwal, P. K., & Bain, P. M. (2019). *Powerful teaching: Unleash the science of learning.* San Francisco, CA: Jossey-Bass.

Agarwal, P. K., Bain, P. M., & Chamberlain, R. W. (2012). The value of applied research: Retrieval practice improves classroom learning and recommendations from a teacher, a principal, and a scientist. *Educational Psychology Review*, *24*, 437-448.

Agarwal, P. K., D'Antonio, L., Roediger, H. L., 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 and Cognition*, *3*, 131-139.

Agarwal, P. K., Finley, J. R., Rose, N. S., & Roediger, H. L. (2017). Benefits from retrieval

practice are greater for students with lower working memory capacity. *Memory*, *25*, 764-771.

Allen, I. E., Seaman, J., Poulin, R., & Straut, T. T. (2016). *Online report card: Tracking online education in the United States*. Babson Survey Research Group. https://files.eric.ed.gov/fulltext/ED572777.pdf

Anderson, C. A., Lindsay, J. J., & Bushman, B. J. (1999). Research in the psychological laboratory: Truth or triviality? *Current Directions in Psychological Science*, *8*, 3-9.

*Atabek Yigit, E., Balkan Kiyici, F., & Çetinkaya, G. (2014). Evaluating the testing effect in the classroom: An effective way to retrieve learned information. *Eurasian Journal of Educational Research*, *54*, 99-116.

Augusteijn, H. E., van Aert, R., & van Assen, M. A. (2019). The effect of publication bias on the Q test and assessment of heterogeneity. *Psychological Methods*, *24*, 116-134.

*Ayyub, A., & Mahboob, U. (2017). Effectiveness of test-enhanced learning (TEL) in lectures for undergraduate medical students. *Pakistan Journal of Medical Sciences*, *33*, 1339.

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

*Batsell, W. R., 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.

Bauer, P. J. (2020). Expanding the reach of Psychological Science [Editorial]. *Psychological Science*, *31*, 3-5.

Becker-Blease, K. A., & Bostwick, K. C. P. (2016). Adaptive quizzing in introductory psychology: Evidence of limited effectiveness. *Scholarship of Teaching and Learning in*

Psychology, 2, 75-86.

*Bekkink, M. O., Donders, R., van Muijen, G. N., & 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.

*Bjork, E. L., Little, J. L., & Storm, B. C. (2014). Multiple-choice testing as a desirable difficulty in the classroom. *Journal of Applied Research in Memory and Cognition*, *3*, 165-170.

Bojnova, E. D., & Oigara, J. N. (2011). Teaching and learning with clickers: are clickers good for students. *Interdisciplinary Journal of E-Learning and Learning Objects*, *7*, 169-184.

Brame, C. J., & Biel, R. (2015). Test-enhanced learning: The potential for testing to promote greater learning in undergraduate science courses. *CBE–Life Sciences Education*, *14*, 1-12.

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

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.

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.

Butler, A. C., & Carpenter, S. K. (2015). Separating myth from reality in education: Introduction to the special issue. *Educational Psychology Review*, *27*, 563-565.

Butler, A. C., Marsh, E. J., Slavinsky, J. P., & Baraniuk, R. G. (2014). Integrating cognitive

science and technology improves learning in a STEM classroom. *Educational Psychology Review*, *26*, 331-340.

Carpenter, S. K., & Agarwal, P. K. (2020). *How to use spaced retrieval practice to boost learning*. http://www.retrievalpractice.org

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.

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.

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.

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

Conway, M. A., Wang, Q., Hanyu, K., & Haque, S. (2005). A cross-cultural investigation of autobiographical memory: On the universality and cultural variation of the reminiscence bump. *Journal of Cross-Cultural Psychology*, *36*, 739-749.

Coyne, J. H., Borg, J. M., DeLuca, J., Glass, L., & Sumowski, J. F. (2015). Retrieval practice as an effective memory strategy in children and adolescents with traumatic brain injury. *Archives of Physical Medicine and Rehabilitation*, *96*, 742-745.

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

Daniel, D. B., & Broida, J. (2004). Using web-based quizzing to improve exam performance: Lessons learned. Teaching of Psychology, 31, 207-208.

de Carvalho Filho, M. K. (2010). Assessing changes in performance and monitoring processes in individual and collaborative tests according to students' metacognitive skills. *European Journal of Cognitive Psychology*, *22*, 1107-1136.

de Lima, N. K., & Jaeger, A. (2020). The effects of prequestions versus postquestions on memory retention in children. *Journal of Applied Research in Memory and Cognition*, *9*, 555-563.

DeLozier, S. J., & Rhodes, M. G. (2017). Flipped classrooms: A review of key ideas and recommendations for practice. *Educational Psychology Review*, *29*, 141-151.

Diaper, G. (1990). The Hawthorne effect: A fresh examination. *Educational Studies*, *16*, 261-267.

*Dirkx, K. J., Kester, L., & Kirschner, P. A. (2014). The testing effect for learning principles and procedures from texts. *The Journal of Educational Research*, *107*, 357-364.

*Dobson, J. L., & Linderholm, T. (2015a). The effect of selected "desirable difficulties" on the ability to recall anatomy information. *Anatomical Sciences Education*, *8*, 395-403.

*Dobson, J. L., & Linderholm, T. (2015b). Self-testing promotes superior retention of anatomy and physiology information. *Advances in Health Sciences Education*, *20*, 149-161.

*Dobson, J. L., Perez, J., & Linderholm, T. (2017). Distributed retrieval practice promotes superior recall of anatomy information. *Anatomical Sciences Education*, *10*, 339-347.

Duchastel, P. C. (1979). Retention of prose materials: The effect of testing. *The Journal of Education Research*, *72*, 299-300.

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.

Dunlosky, J., & Rawson, K. A. (Eds.) (2019). *The Cambridge Handbook of Cognition and Education*. Cambridge University Press.

Eisenkraemer, R. E., Jaeger, A., & Stein, L. M. (2013). A systematic review of the testing effect in learning. *Paidéia*, *23*, 397-406.

Ekuni, R., de Souza, B. M. N., Agarwal, P. K., & Pompeia, S. (2020). A conceptual replication of survey research on study strategies in a diverse, non-WEIRD student population. *Scholarship of Teaching and Learning in Psychology*. https://doi.org/10.1037/stl0000191

Fazio, L. K., & Marsh, E. J. (2019). Retrieval-based learning in children. *Current Directions in Psychological Science*, *28*, 111-116.

Ferguson, C. J., & Heene, M. (2012). A vast graveyard of undead theories: Publication bias and psychological science's aversion to the null. *Perspectives on Psychological Science*, *7*, 555-561.

Francis, A. P., Wieth, M. B., Zabel, K. L., & Carr, T. H. (2020). A classroom study on the role of prior knowledge and retrieval tool in the testing effect. *Psychology Learning & Teaching*. Advance online publication. doi:10.1177/1475725720924872

*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.

*Freda, N. M., & Lipp, M. J. (2016). Test-enhanced learning in competence-based predoctoral orthodontics: a four-year study. *Journal of Dental Education*, *80*, 348-354.

Gates, A. I. (1917). Recitation as a factor in memorizing. Archives of Psychology, 6(40).

Glover, J. A. (1989). The ‘‘testing’’ phenomenon: Not gone but nearly forgotten. Journal of Educational Psychology, 81, 392-399.

*Goossens, N. A., Camp, G., Verkoeijen, P. P., 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.

*Graham, R. B. (1999). Unannounced quizzes raise test scores selectively for mid-range students. *Teaching of Psychology*, *26*, 271-273.

Green, M. L., Moeller, J. J., & Spak, J. M. (2018). Test-enhanced learning in health professions education: A systematic review. *Medical Teacher*, *40*, 337-350.

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.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*, 1-75.

Jaeger, A., Eisenkraemer, R. E., & Stein, L. M. (2015). Test-enhanced learning in third-grade children. *Educational Psychology*, *35*, 513-521.

*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.

Kang, S. H. K., & Pashler, H. (2014). Is the benefit of retrieval practice modulated by motivation? *Journal of Applied Research in Memory and Cognition*, *3*, 183-188.

Karpicke, J. D. (2012). Retrieval-based learning: Active retrieval promotes meaningful learning. *Current Directions in Psychological Science*, *21*, 157-163.

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*, *7*, 1-9.

*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 and Cognition*, *3*, 198-206.

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

Karpicke, J. D., & Grimaldi, P. J. (2012). Retrieval-based learning: A perspective for enhancing meaningful learning. *Educational Psychology Review*, *24*, 401-418.

Kelley, K. (2007a). Methods for the Behavioral, Educational, and Social Sciences: An R Package. *Behavior Research Methods*, *39*, 979-984.

Kelley, K. (2007b). Confidence intervals for standardized effect sizes: Theory, application, and implementation. *Journal of Statistical Software*, *20*, 1-24.

Kelley, K. (2017). MBESS (Version 4.0.0 and higher) [computer software and manual]. http://cran.r-project.org

*Khanna, M. M. (2015). Ungraded pop quizzes: Test-enhanced learning without all the anxiety. *Teaching of Psychology*, *42*, 174-178.

*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.

Kibble, J. (2007). Use of unsupervised online quizzes as formative assessment in a medical physiology course: effects of incentives on student participation and performance. *Advances in Physiology Education*, *31*, 253-260.

Kornell, N., Rabelo, V. C., and Klein, P. C. (2012). Tests enhance learning—Compared to what?

Journal of Applied Research in Memory & Cognition, *1*, 257-259.

*Kromann, C. B., Jensen, M. L., & Ringsted, C. (2009). The effect of testing on skills learning. *Medical Education*, *43*, 21-27.

*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.

*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.

Kulik, J. A., & Kulik, C. C. (1988). Timing of feedback and verbal learning. *Review of Educational Research*, *58*, 79-97.

Lakens, D., Hilgard, J., & Staaks, J. (2016). On the reproducibility of meta-analyses: Six practical recommendations. *BMC Psychology*, *4*, 24.

Lan, X., Legare, C. H., Ponitz, C. C., Li, S., & Morrison, F. J. (2011). Investigating the links between the subcomponents of executive function and academic achievement: A cross-cultural analysis of Chinese and American preschoolers. *Journal of Experimental Child Psychology*, *108*, 677-692.

*Larsen, D. P., Butler, A. C., & Roediger, H. L. (2009). Repeated testing improves long-term retention relative to repeated study: a randomised controlled trial. *Medical Education*, *43*, 1174-1181.

*Larsen, D. P., Butler, A. C., & Roediger, H. L. (2013a). Comparative effects of test-enhanced learning and self-explanation on long-term retention. *Medical Education*, *47*, 674-682.

*Larsen, D. P., Butler, A. C., Lawson, A. L., & Roediger, H. L. (2013b). 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.

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

Lipko-Speed, A., Dunlosky, J., and 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.

*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.

Markus, H. R., & Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review*, *98*, 224.

*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.

*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.

*McDermott, K. B., Agarwal, P. K., D'Antonio, L., Roediger, H. L., & 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.

McLaughlin, K., & Coderre, S. (2015). The potential and conditional benefits of retrieval practice on learning. *Advances in Health Science Education*, *20*, 321-324.

McShane, B. B., & Böckenholdt, U. (2020). Enriching meta-analytic models of summary data: A

thought experiment and case study. *Advances in Methods and Practices in Psychological Science*, *3*, 81-93.

Metcalfe, J., Kornell, N., & Finn, B. (2009). Delayed versus immediate feedback in children's and adults' vocabulary learning. *Memory & Cognition*, *37*, 1077-1087.

*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.

Minear, M., Coane, J. H., Boland, S. C., Cooney, L. H., & Albat, M. (2018). The benefits of retrieval practice depend on item difficulty and intelligence. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*, 1474-1486.

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*, 1-16.

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.

Mullet, H. G., Butler, A. C., Verdin, B., von Borries, R., & Marsh, E. J. (2014). Delaying feedback promotes transfer of knowledge despite student preferences to receive feedback immediately. *Journal of Applied Research in Memory and Cognition*, *3*, 222-229.

Myers, G. C. (1914). Recall in relation to retention. *Journal of Educational Psychology*, *5*, 119-130.

*Narloch, R., Garbin, C. P., & Turnage, K. D. (2006). Benefits of prelecture quizzes. *Teaching of Psychology*, *33*, 109-112.

Nguyen, K., & McDaniel, M. A. (2015). Using quizzing to assist student learning in the classroom: The good, the bad, and the ugly. *Teaching of Psychology*, *42*, 87-92.

Niedermeyer, F. C., & Sullivan, H. J. (1972). Differential effects of individual and group testing strategies in an objectives-based instructional program. *Journal of Educational Measurement*, *9*, 199-204.

Nunes, L. D., & Karpicke, J. D. (2015). Retrieval-based learning: Research at the interface between cognitive science and education. In R. A. Scott & S. M. Kosslyn (Eds.), *Emerging Trends in the Social and Behavioral Sciences* (pp. 1-16). John Wiley & Sons, Inc.

O'Donnell, C. L. (2008). Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K–12 curriculum intervention research. *Review of Educational Research*, *78*, 33-84.

OECD (2020). Education at a glance: Educational attainment and labour-force status. https://data.oecd.org/eduatt/population-with-tertiary-education.htm

Open Science Framework. (2020). Center for Open Science. http://cos.io/osf

Pachai, A. A., Acai, A., LoGiudice, A. B., & Kim, J. A. (2016). The mind that wanders: Challenges and potential benefits of mind wandering in education. *Scholarship of Teaching and Learning in Psychology*, *2*, 134-146.

Pan. S. C., & Agarwal, P. K. (2020). *Retrieval practice and transfer of learning: Fostering students' application of knowledge*. http://www.retrievalpractice.org

Pan, S. C., and Rickard, T. C. (2018). Transfer of test-enhanced learning: Meta-analytic review and synthesis. *Psychonomic Bulletin*, *144*, 710-756.

Paradis, E., & Sutkin, G. (2017). Beyond a good story: from Hawthorne Effect to reactivity in health professions education research. *Medical Education*, *51*, 31-39.

Pyc, M. A., Agarwal, P. K., & Roediger, H. L. (2014). Test-enhanced learning. Chapter in V. A.

Benassi, C. E. Overson, & C. M. Hakala (Eds.), *Applying science of learning in education: Infusing psychological science into the curriculum.* APA Society for the Teaching of Psychology.

Racsmány, M., Szőllősi, A., & Bencze, D. (2018). Retrieval practice makes procedure from remembering: An automatization account of the testing effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*, 157-166.

Rad, M. S., Martingano, A. J., & Ginges, J. (2018). Toward a psychology of Homo sapiens: Making psychological science more representative of the human population. *Proceedings of the National Academy of Sciences*, *115*, 11401-11405.

Rawson, K. A., Dunlosky, J., & Sciartelli, S. M. (2013). The power of successive relearning: Improving performance on course exams and long-term retention. *Educational Psychology Review*, *25*, 523-548.

*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.

Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, *15*, 20-27.

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.

Roediger, H. L., & Karpicke, J. D. (2006b). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*, 249-255.

Rohrer, D., Dedrick, R. F., Hartwig, M. K., & Cheung, C.-N. (2020). A randomized controlled trial of interleaved mathematics practice. *Journal of Educational Psychology*, *112*, 40-52.

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.

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

Schwieren, J., Barenberg, J., & Dutke, S. (2017). The testing effect in the psychology classroom: A meta-analytic perspective. *Psychology Learning & Teaching*, *16*, 179-196.

Segall, M. H., Campbell, D. T. & Herskovits, M. J. (1966). *The influence of culture on visual perception*. Indianapolis: Bobbs-Merrill.

Sennhenn-Kirchner, S., Goerlich, Y., Kirchner, B., Notbohm, M., Schiekirka, S., Simmenroth, A., & Raupach, T. (2016). The effect of repeated testing vs repeated practice on skills learning in undergraduate dental education. *European Journal of Dental Education*, *22*, e42-e47.

*Shapiro, A. M., & Gordon, L. T. (2012). A controlled study of clicker-assisted memory enhancement in college classrooms. *Applied Cognitive Psychology*, *26*, 635-643.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22, 1359-1366.

*Son, J. Y., & Rivas, M. J. (2016). Designing clicker questions to stimulate transfer. *Scholarship of Teaching and Learning in Psychology*, *2*, 193-207.

Sotola, L. K., & Crede, M. (2020). Regarding class quizzes: A meta-analytic synthesis of studies on the relationship between frequent low-stakes testing and class performance. *Educational Psychology Review*. doi.org/10.1007/s10648-020-09563-9

Spitzer, H. F. (1939). Studies in retention. *Journal of Educational Psychology*, *30*, 641-656.

*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.

Tipton, E., Pustejovsky, J. E., & Ahmadi, H. (2019). Current practices in meta-regression in psychology, education, and medicine. *Research Synthesis Methods*, *10*, 180-194.

*Tu, Y. C., Lin, Y. J., Lee, J. W., & Fan, L. W. (2017). Effects of didactic instruction and test-enhanced learning in a nursing review course. *Journal of Nursing Education*, *56*, 683-687.

Tweed, R. G., & Lehman, D. R. (2002). Learning considered within a cultural context: Confucian and Socratic approaches. *American Psychologist*, *57*, 89-99.

van Egmond, M. C., Kühnen, U., & Li, J. (2013). Mind and virtue: The meaning of learning, a matter of culture? *Learning, Culture and Social Interaction*, *2*, 208-216.

Viveiros, J., Sethares, K., & Shapiro, A. (2017). Repeated recall as an intervention to improve memory performance in heart failure patients. *European Journal of Cardiovascular Nursing*, *16*, 724-732.

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.

Weinstein, Y., Madan, C. R., & Sumeracki, M. A. (2018). Teaching the science of learning. *Cognitive Research: Principles and Implications*, *3*, 2.

Weinstein, Y., Nunes, L. K., & Karpicke, J. D. (2016). On the placement of practice questions during study. *Journal of Experimental Psychology: Applied*, *22*, 72-84.

Wiklund-Hörnqvist, C., Jonsson, B., and Nyberg, L. (2014). Strengthening concept learning by

repeated testing. *Scandinavian Journal of Psychology*, *55*, 10-16.

Zotero. (2020). Corporation for Digital Scholarship. http://www.zotero.org

Abbreviated Citation	Retrieval Practice Intervention	Comparison Condition(s)	Effect Sizes (Cohen's *d*) Asterisk (*) indicates insufficient data to calculate effect size	Education Level	Overall Content Area	Course Topic	Experimental Design (within- or between-students) Asterisk (*) indicates random assignment	Sample Size in Final Analyses
Agarwal, 2019, Experiment 3	Quizzes with mixed (fact and higher order) questions Quizzes with higher order questions	Questions on the final assessment only	1.45 [1.15, 1.75] (mixed quizzes vs. not quizzed; fact test) 1.30 [1.01, 1.59] (mixed quizzes vs. not quizzed; higher order test) 0.84 [0.59, 1.08] (higher order quizzes vs. not quizzed; higher order test)	Middle School	History	World History	Within	88
Atabek Yigit et al., 2014	Quizzes with feedback Quizzes without feedback	Studied a review sheet No review sheet	Numerical benefit from retrieval practice for all comparisons*	Undergraduate	Science	Introduction to Chemistry	Within	98
Ayyub & Mahboob, 2017	Quizzes	Lessons without quizzes	Numerical benefit from retrieval practice*	Medical School	Science	Endocrinology	Between*	84
Batsell et al., 2017	Textbook reading with in-class quizzes	Textbook reading without quizzes	Numerical benefit from retrieval practice*	Undergraduate	Psychology	Introduction to Psychology	Between	64
Bekkink et al., 2012	Quizzes	Lessons without quizzes	0.24 [0.03, 0.45]	Medical School	Science	Introduction to Pathology	Between*	404
Bjork et al., 2014	Quizzes with identical questions repeated on the final assessment Quizzes with conceptually related questions on the final assessment	Questions on the final assessment only	0.94 [0.81, 1.06] (identical-repeat quizzes vs. not quizzed) 0.78 [0.66, 0.89] (conceptual-repeat quizzes vs. not quizzed)	Undergraduate	Psychology	Research Methods in Psychology	Within	372
Dirkx et al., 2014	Study-test-study-test (STST)	Study-study-study-study (SSSS)	1.42 [0.70, 2.13] (final test with factual questions) 0.96 [0.28, 1.63] (final test with application questions)	High School	Mathematics	Statistics	Between*	38
Dobson & Linderholm, 2015(a)	Read-quiz-read-quiz (RTRT)	Read-generate-read-generate (RGRG) Re-read (RRRR)	Numerical benefit from retrieval practice for all comparisons*	Undergraduate	Science	Anatomy and Physiology	Within	66
Dobson & Linderholm, 2015(b)	Read-quiz-read (RTR)	Read-read while taking notes (R-R+N) Re-read (RRR)	Numerical benefit from retrieval practice for all comparisons*	Undergraduate	Science	Anatomy and Physiology	Within	125
Dobson et al., 2017	Free recall over 7 days Free recall over 2 days Free recall in a single session	Re-study over 7 days Re-study over 2 days	Numerical benefit from retrieval practice for all comparisons*	Undergraduate	Science	Anatomy and Physiology	Within	60

Abbreviated Citation	Retrieval Practice Intervention	Comparison Condition(s)	Effect Sizes (Cohen's *d*) Asterisk (*) indicates insufficient data to calculate effect size	Education Level	Overall Content Area	Course Topic	Experimental Design (within- or between-students) Asterisk (*) indicates random assignment	Sample Size in Final Analyses
Foss & Pirozzolo, 2017, Experiment 1	4 midterm exams and 4 pop quizzes	2 midterm exams and 2 pop quizzes	Numerical benefit from retrieval practice*	Undergraduate	Psychology	Research Methods in Psychology	Between	159
Foss & Pirozzolo, 2017, Experiment 2	8 course exams	2 course exams	Numerical benefit from retrieval practice*	Undergraduate	Psychology	Research Methods in Psychology	Between	70
Foss & Pirozzolo, 2017, Experiment 3	8 course exams and additional quizzes	2 course exams	0.25 [-0.23, 0.73]	Undergraduate	Psychology	Research Methods in Psychology	Between	117
Foss & Pirozzolo, 2017, Experiment 4	6 quizzes	3 quizzes	0.21 [-0.12, 0.55]	Undergraduate	Psychology	Research Methods in Psychology	Between	214
Freda & Lipp, 2016	Quizzes	Lessons without quizzes	Numerical benefit from retrieval practice*	Medical School	Skills-Based	Dental Diagnosis and Treatment	Between	338
Goossens et al., 2016	Cued recall	Copying definitions	Numerical benefit from copying definitions*	Elementary School	Spelling and Vocabulary	Vocabulary	Within	129
Graham, 1999	Textbook reading with in-class quizzes	Textbook reading without quizzes	0.27 [0.09, 0.44]	Undergraduate	Psychology	Neuropsychology and Psychology of Learning	Within	Not Reported
Jones et al., 2016, Experiment 1	Quizzes	Rainbow writing	0.60 [0.03, 1.18]	Elementary School	Spelling and Vocabulary	Spelling	Within	14
Jones et al., 2016, Experiment 2	Quizzes	Rainbow writing	0.69 [0.14, 1.23]	Elementary School	Spelling and Vocabulary	Spelling	Within	16
Jones et al., 2016, Experiment 3	Quizzes	Rainbow writing	1.61 [0.73, 2.48]	Elementary School	Spelling and Vocabulary	Spelling	Within	12

Abbreviated Citation	Retrieval Practice Intervention	Comparison Condition(s)	Effect Sizes (Cohen's *d*) Asterisk (*) indicates insufficient data to calculate effect size	Education Level	Overall Content Area	Course Topic	Experimental Design (within- or between-students) Asterisk (*) indicates random assignment	Sample Size in Final Analyses
Karpicke et al., 2014, Experiment 3	Cued recall using a concept map	Re-study	0.42 [0.20, 0.65]	Elementary School	Science and History	Earth Science and U.S. History	Within	85
Khanna & Cortese, 2016	Graded quizzes Ungraded quizzes	Lessons without quizzes	0.36 [0.03, 0.69] (ungraded quizzes vs. no quizzes) 0.28 [-0.05, 0.61] (graded quizzes vs. no quizzes)	Undergraduate	Psychology	Introduction to Psychology, Cognitive Psychology, and Research Methods	Between*	277
Khanna, 2015	Graded quizzes Ungraded quizzes	Lessons without quizzes	0.32 [-0.09, 0.72] (ungraded quizzes vs. no quizzes) -0.27 [-0.68, 0.14] (graded quizzes vs. no quizzes)	Undergraduate	Psychology	Introduction to Psychology	Between*	137
Kromann et al., 2009	Scenario-based quiz	Scenario lectures without quizzes	Numerical benefit from retrieval practice*	Medical School	Skills-Based	CPR Skills	Between*	81
Kromann et al., 2010	Scenario-based quiz	Scenario lectures without quizzes	0.39 [-0.02, 0.79]	Medical School	Skills-Based	CPR Skills	Between*	89
Kromann et al., 2011	Scenario-based quiz	Scenario lectures without quizzes	0.55 [0.21, 0.89]	Medical School	Skills-Based	CPR Skills	Between*	138
Larsen et al., 2009	Quizzes	Studied a review sheet	0.62 [0.27, 0.96]	Medical School	Science	Neurology	Within	40
Larsen et al., 2013(a)	Quizzes with self-explanations Quizzes without self-explanations	Studied a review sheet with self-explanations Studied a review sheet without self-explanations	0.70 [0.38, 1.01] (quizzes with self-explanation vs. study with self-explanation) 0.48 [0.18, 0.77] (quizzes without self-explanation vs. study with self-explanation) Numerical benefit from retrieval practice for remaining comparisons*	Medical School	Science	Neurology	Within	49
Larsen et al., 2013(b)	Quizzes with standardized patients (SP) Quizzes with written questions	Studied a review sheet	0.80 [0.44, 1.15] (SP quiz vs. study; final SP test) 0.69 [0.35, 1.03] (SP quiz vs. study; final written test) 0.33 [0.02, 0.65] (written quiz vs. study; final SP test) 0.72 [0.37, 1.06] (written quiz vs. study; final written test)	Medical School	Science	Neurology	Within	41
Leeming, 2002	Quizzes every class	4 course exams	0.47 [0.01, 0.94]	Undergraduate	Psychology	Introduction to Psychology and Psychology of Learning	Between	192

Abbreviated Citation	Retrieval Practice Intervention	Comparison Condition(s)	Effect Sizes (Cohen's *d*) Asterisk (*) indicates insufficient data to calculate effect size	Education Level	Overall Content Area	Course Topic	Experimental Design (within- or between-students) Asterisk (*) indicates random assignment	Sample Size in Final Analyses
Lyle & Crawford, 2011	Quizzes	Lessons without quizzes	Numerical benefit from retrieval practice*	Undergraduate	Mathematics	Statistics for Psychology	Between	144
McDaniel et al., 2011, Experiment 1	Quizzes	Questions on the final assessment only	Numerical benefit from retrieval practice*	Middle School	Science	Biology	Within	92
McDaniel et al., 2011, Experiment 2a	Immediate quizzes Delayed quizzes	Questions on the final assessment only	0.48 [0.22, 0.74] (immediate quizzes vs. not quizzed) 0.89 [0.60, 1.18] (delayed quizzes vs. not quizzed)	Middle School	Science	Astronomy	Within	65
McDaniel et al., 2011, Experiment 2b	Immediate quizzes Delayed quizzes	Questions on the final assessment only	0.53 [0.25, 0.82] (immediate quizzes vs. not quizzed) 0.58 [0.29, 0.87] (delayed quizzes vs. not quizzed)	Middle School	Science	Biology and Chemistry	Within	54
McDaniel et al., 2013, Experiment 1	Quizzes with definition questions Quizzes with key term questions	Questions on the final assessment only	Numerical benefit from retrieval practice for all comparisons*	Middle School	Science	Biology and Physics	Within	61
McDaniel et al., 2013, Experiment 2	Quizzes with definition questions Quizzes with key term questions	Questions on the final assessment only	Numerical benefit from retrieval practice for all comparisons*	Middle School	Science	Earth Science	Within	90
McDermott et al., 2014, Experiment 1a	Quizzes with multiple-choice questions Quizzes with short answer questions	Questions on the final assessment only	0.87 [0.52, 1.21] (multiple-choice quizzes vs. not quizzed; final test with short answer questions) 0.48 [0.17, 0.79] (short answer quizzes vs. not quizzed; final test with short answer questions) Numerical benefit from retrieval practice (multiple-choice quizzes vs. not quizzed; final test with multiple-choice questions) Equivalent means for retrieval practice and comparison conditions (short answer quizzes vs. not quizzed; final test with multiple-choice questions)	Middle School	Science	Earth Science	Within	45
McDermott et al., 2014, Experiment 1b	Quizzes with multiple-choice questions Quizzes with short answer questions	Questions on the final assessment only	Numerical benefit from retrieval practice for all comparisons*	Middle School	Science	Biology	Within	45
McDermott et al., 2014, Experiment 2	Quizzes	Re-study Questions on the final assessment only	1.58 [1.19, 1.97] (quizzes vs. re-study) 2.19 [1.71, 2.66] (quizzes vs. not quizzed)	Middle School	Science	Biology and Physics	Within	59
McDermott et al., 2014, Experiment 3	Quizzes with multiple-choice questions Quizzes with short answer questions	Re-study Questions on the final assessment only	Numerical benefit from retrieval practice for all comparisons*	Middle School	Science	Biology and Earth Science	Within	60

Abbreviated Citation	Retrieval Practice Intervention	Comparison Condition(s)	Effect Sizes (Cohen's *d*) Asterisk (*) indicates insufficient data to calculate effect size	Education Level	Overall Content Area	Course Topic	Experimental Design (within- or between-students) Asterisk (*) indicates random assignment	Sample Size in Final Analyses
McDermott et al., 2014, Experiment 4	Quizzes with multiple-choice questions Quizzes with short answer questions	Questions on the final assessment only	Numerical benefit from retrieval practice for all comparisons*	High School	History	U.S. and World History	Within	40
Michaels, 2017	Quizzes (freshmen and sophomores) Quizzes (juniors and seniors)	Lessons without quizzes	0.44 [0.07, 0.82] (freshmen and sophomores) -0.30 [-1.05, 0.46] (juniors and seniors)	Undergraduate	Psychology	Introduction to Psychology	Between	139
Narloch et al., 2006	Pre-lecture quizzes with matching questions Pre-lecture quizzes with fill-in-the-blank questions	Lessons without quizzes	Numerical benefit from retrieval practice for all comparisons*	Undergraduate	Psychology	Sensation & Perception	Between	109
Roediger et al., 2011, Experiment 1	Quizzes	Questions on the final assessment only	Numerical benefit from retrieval practice*	Middle School	History	World History	Within	36
Roediger et al., 2011, Experiment 2	Quizzes	Re-study Questions on the final assessment only	0.83 [0.54, 1.12] (quizzes vs. re-study) 0.96 [0.66, 1.25] (quizzes vs. not quizzed)	Middle School	History	World History	Within	63
Saville et al., 2012	Quizzes	Lessons without quizzes	No numerical benefit from retrieval practice*	Undergraduate	Psychology	Psychology of Learning	Within	58
Shapiro & Gordon, 2012	Quizzes	Questions on the final assessment only	0.38 [0.25, 0.52]	Undergraduate	Psychology	Introduction to Psychology	Within	226
Son & Rivas, 2016	Quizzes	Studied a review sheet	0.45 [0.17, 0.73] (final test with repeated questions) 0.14 [-0.14, 0.42] (final test with transfer questions)	Undergraduate	Psychology	Developmental Psychology	Between	209
Stenlund et al., 2017	Quizzes	Discussion with feedback Discussion without feedback	1.54 [0.99, 2.08] (quizzes vs. discussion with feedback; final test with complex questions) 0.69 [0.20, 1.18] (quizzes vs. discussion with feedback; final test with factual questions) 1.06 [0.54, 1.59] (quizzes vs. discussion without feedback; final test with complex questions) 0.72 [0.21, 1.23] (quizzes vs. discussion without feedback; final test with factual questions)	High School	Psychology	Psychology of Emotions	Between*	98
Tu et al., 2017	Quizzes	Lessons without quizzes	-0.24 [-0.67, 0.19]	Medical School	Skills-Based	Fundamentals of Nursing	Between	84

Abbreviated Citation	Conducted in the United States	Retrieval Practice Timing	Delay Between the Last Retrieval Opportunity and the Final Test	Retrieval Practice Format	Feedback After Retrieval Practice	Final Test Format Asterisk (*) indicates rephrased final test questions	Final Test Peformance Counted Toward Students' Grades
Agarwal, 2019, Experiment 3	Yes	Three times within 1 week	2 days	Multiple-choice	Immediate	Multiple-choice*	No
Atabek Yigit et al., 2014	No (Turkey)	Once within a single session	1 day	Multiple-choice and matching	Immediate	Multiple-choice and matching	No
Ayyub & Mahboob, 2017	No (Pakistan)	Once per week for 4 weeks	1 week	Multiple-choice	Not reported	Multiple-choice*	Yes
Batsell et al., 2017	Yes	Twice per week throughout the semester (timing not specified)	Three exams throughout the semester (timing not specified)	Not Reported	Immediate	Multiple-choice*	Yes
Bekkink et al., 2012	No (The Netherlands)	1-2 times per week for 4 weeks	3 days	Multiple-choice and short answer	None	Multiple-choice*	Yes
Bjork et al., 2014	Yes	Once every 2 weeks for 10 weeks	End of the 10-week course	Multiple-choice	Delayed	Multiple-choice*	Yes
Dirkx et al., 2014	No (The Netherlands)	Once within a single session	1 week	Short answer	None	Short answer	No
Dobson & Linderholm, 2015(a)	Yes	Twice within a single session	Immediate	Free recall	None	Free recall	No
Dobson & Linderholm, 2015(b)	Yes	Once within a single session	Immediate	Free recall	None	Multiple-choice	No
Dobson et al., 2017	Yes	Six times within 7 days	Immediate	Free recall	None	Free recall	No

Abbreviated Citation	Conducted in the United States	Retrieval Practice Timing	Delay Between the Last Retrieval Opportunity and the Final Test	Retrieval Practice Format	Feedback After Retrieval Practice	Final Test Format Asterisk (*) indicates rephrased final test questions	Final Test Peformance Counted Toward Students' Grades
Foss & Pirozzolo, 2017, Experiment 1	Yes	Once every 2 weeks for 15 weeks	End of the 15-week course	Multiple-choice and short answer	None	Multiple-choice and short answer	Yes
Foss & Pirozzolo, 2017, Experiment 2	Yes	Once every 2 weeks for 15 weeks	End of the 15-week course	Multiple-choice and short answer	None	Multiple-choice and short answer	Yes
Foss & Pirozzolo, 2017, Experiment 3	Yes	Once every 2 weeks for 15 weeks	End of the 15-week course	Multiple-choice and short answer	None	Multiple-choice and short answer*	Yes
Foss & Pirozzolo, 2017, Experiment 4	Yes	Once every 2 weeks for 15 weeks	7-10 days	Short answer	None	Short answer*	Yes
Freda & Lipp, 2016	Yes	Once per week for 6 weeks	End of the 6-week course	Simulated diagnosis	Delayed	Simulated diagnosis*	Yes
Goossens et al., 2016	No (The Netherlands)	Six times total for 1-2 weeks	1-3 days	Cued recall	Immediate	Cued recall	Not reported
Graham, 1999	Yes	10 times throughout the semester (timing not specified)	End of the semester	Multiple-choice and fill-in-the-blank	Not Reported	Multiple-choice and fill-in-the-blank*	Yes
Jones et al., 2016, Experiment 1	Yes	Twice within 8 days	1 day	Free recall	Immediate	Free recall	Yes
Jones et al., 2016, Experiment 2	Yes	Twice within 8 days	1 day	Free recall	Immediate	Free recall	Yes
Jones et al., 2016, Experiment 3	Yes	Twice within 8 days	1 day	Free recall	Immediate	Free recall	Yes

Abbreviated Citation	Conducted in the United States	Retrieval Practice Timing	Delay Between the Last Retrieval Opportunity and the Final Test	Retrieval Practice Format	Feedback After Retrieval Practice	Final Test Format Asterisk (*) indicates rephrased final test questions	Final Test Peformance Counted Toward Students' Grades
Karpicke et al., 2014, Experiment 3	Yes	Once within a single session	Immediate	Short answer	None	Free recall	No
Khanna & Cortese, 2016	Yes	Six times throughout the semester (timing not specified)	End of the semester	Multiple-choice	Immediate	Multiple-choice*	Yes
Khanna, 2015	Yes	Six times throughout the semester (timing not specified)	End of the semester	Multiple-choice	Immediate	Multiple-choice*	Yes
Kromann et al., 2009	No (Denmark)	Once within a single session	2 weeks	Simulated diagnosis	Immediate	Simulated diagnosis	No
Kromann et al., 2010	No (Denmark)	Once within a single session	6 months	Simulated diagnosis	Immediate	Simulated diagnosis	No
Kromann et al., 2011	No (Denmark)	Once within a single session	2 weeks	Simulated diagnosis	Immediate	Simulated diagnosis	No
Larsen et al., 2009	Yes	Three times within a single session	6 months	Short answer	Immediate	Short answer	No
Larsen et al., 2013(a)	Yes	Four times within 2 days	6 months	Short answer	Immediate	Essay*	No
Larsen et al., 2013(b)	Yes	Four times within a single session	6 months	Simulated diagnosis and a short answer quiz	Immediate	Simulated diagnosis and a short answer test*	No
Leeming, 2002	Yes	5 times per week for 5 weeks (summer semester) or 2 times per week for 12 weeks (regular semester)	6 weeks	Short answer	Immediate	Short answer, multiple-choice, and fill-in-the-blank	No

Abbreviated Citation	Conducted in the United States	Retrieval Practice Timing	Delay Between the Last Retrieval Opportunity and the Final Test	Retrieval Practice Format	Feedback After Retrieval Practice	Final Test Format Asterisk (*) indicates rephrased final test questions	Final Test Peformance Counted Toward Students' Grades
Lyle & Crawford, 2011	Yes	Once per week for 21 weeks	Four exams throughout the semester (timing not specified)	Fill-in-the-blank	Delayed	Multiple-choice*	Yes
McDaniel et al., 2011, Experiment 1	Yes	Three times within 20 days	1 day	Multiple-choice	Immediate	Multiple-choice	Yes
McDaniel et al., 2011, Experiment 2a	Yes	1-3 times within 3-8 days	1 day	Multiple-choice	Immediate	Multiple-choice	Yes
McDaniel et al., 2011, Experiment 2b	Yes	1-3 times within 1-4 days	1 day	Multiple-choice	Immediate	Multiple-choice	Yes
McDaniel et al., 2013, Experiment 1	Yes	Three times within 11 days	1 day	Multiple-choice	Immediate	Multiple-choice*	Yes
McDaniel et al., 2013, Experiment 2	Yes	Three times within 16 days	1 day	Multiple-choice	Immediate	Multiple-choice*	Yes
McDermott et al., 2014, Experiment 1a	Yes	Three times within 22 days	1 day	Multiple-choice and short answer	Immediate	Multiple-choice and short answer	Yes
McDermott et al., 2014, Experiment 1b	Yes	Three times within 7 days	1 day	Multiple-choice and short answer	Immediate	Multiple-choice and short answer	Yes
McDermott et al., 2014, Experiment 2	Yes	Three times within 7 days	1 day	Short answer	Immediate	Short answer	Yes
McDermott et al., 2014, Experiment 3	Yes	Twice within 7 days	2-3 days	Multiple-choice and short answer	Immediate	Multiple-choice and short answer*	Yes

Abbreviated Citation	Conducted in the United States	Retrieval Practice Timing	Delay Between the Last Retrieval Opportunity and the Final Test	Retrieval Practice Format	Feedback After Retrieval Practice	Final Test Format Asterisk (*) indicates rephrased final test questions	Final Test Peformance Counted Toward Students' Grades
McDermott et al., 2014, Experiment 4	Yes	Twice within 8 days	1-2 days	Multiple-choice and short answer	Immediate	Multiple-choice and short answer*	Yes
Michaels, 2017	Yes	Once per week for 16 weeks	Three exams throughout the semester (timing not specified)	Multiple-choice and diagram labeling	Immediate	Multiple-choice	No
Narloch et al., 2006	Yes	Multiple times throughout the semester (timing not specified)	Exams at the end of each chapter (timing not specified)	Fill-in-the-blank and matching	Delayed	Multiple-choice and essay*	Yes
Roediger et al., 2011, Experiment 1	Yes	Three times within 13 days	2 days	Multiple-choice	Immediate	Free recall	Yes
Roediger et al., 2011, Experiment 2	Yes	Three times within 11 days	2 days	Multiple-choice	Immediate	Multiple-choice	Yes
Saville et al., 2012	Yes	Nine times throughout the semester (timing not specified)	2 days	Multiple-choice	Immediate	Multiple-choice and short answer*	Yes
Shapiro & Gordon, 2012	Yes	Three times per week for 15 weeks	Four exams throughout the semester (timing not specified)	Multiple-choice	Immediate	Multiple-choice*	Yes
Son & Rivas, 2016	Yes	1-3 times per week for 10 weeks	1 week	Multiple-choice	Immediate	Multiple-choice*	Yes
Stenlund et al., 2017	No (Sweden)	Twice per day for 2 days	Three exams over 4 weeks	Short answer	Immediate	Short answer	Not reported
Tu et al., 2017	No (Taiwan)	Every class meeting throughout the semester (timing not specified)	End of the semester	Multiple-choice	Immediate	Multiple-choice	Yes