frontiers in Education

REVIEW
published: 08 February 2019
doi: 10.3389/feduc.2019.00005

Retrieval Practice in Classroom Settings: A Review of Applied Research

*Bruna Fernanda Tolentino Moreira, Tatiana Salazar Silva Pinto, Daniela Siqueira Veloso Starling and Antônio Jaeger**

Department of Psychology, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

Tests have been vastly used for the assessment of learning in educational contexts. Recently, however, a growing body of research has shown that the practice of remembering previously studied information (i.e., retrieval practice) is more advantageous for long-term retention than restudying that same information; a phenomenon often termed "testing effect." The question remains, however, whether such practice can be useful to improve learning in actual educational contexts, and whether in these contexts specific types of tests are particularly beneficial. We addressed these issues by reviewing studies that investigated the use of retrieval practice as a learning strategy in actual educational contexts. The studies reviewed here adopted from free-recall to multiple-choice tests, and involved from elementary school children to medical school students. In general, their results are favorable to the use of retrieval practice in classroom settings, regardless of whether feedback is provided or not. Importantly, however, the majority of the reviewed studies compared retrieval practice to repeated study or to "no-activity." The results of the studies comparing retrieval practice to alternative control conditions were less conclusive, and a subset of them found no advantage for tests. These findings raise the question whether retrieval practice is more beneficial than alternative learning strategies, especially learning strategies and activities already adopted in classroom settings (e.g., concept mapping). Thus, even though retrieval practice emerges as a promising strategy to improve learning in classroom environments, there is not enough evidence available at this moment to determine whether it is as beneficial as alternative learning activities frequently adopted in classroom settings.

Keywords: tests, testing effect, retrieval practice, test-enhanced learning, classroom

OPEN ACCESS

Edited by:
Asimina M. Ralli, National and Kapodistrian University of Athens, Greece

Reviewed by:
Panagiota Dimitropoulou, University of Crete, Greece
Evangelia Karagiannopoulou, University of Ioannina, Greece

****Correspondence:***
Antônio Jaeger
antonio.jaeger@gmail.com

Specialty section:
This article was submitted to Educational Psychology, a section of the journal Frontiers in Education

Received: *10 July 2018*
Accepted: *22 January 2019*
Published: *08 February 2019*

Citation:
Moreira BFT, Pinto TSS, Starling DSV and Jaeger A (2019) Retrieval Practice in Classroom Settings: A Review of Applied Research. Front. Educ. 4:5. doi: 10.3389/feduc.2019.00005

INTRODUCTION

In educational settings, tests are widely used to assess learning. A growing body of research, however, has shown that beyond an assessment tool, tests can also be an effective method to increase long-term retention of studied materials (Bjork, 1988; Roediger and Karpicke, 2006a; Karpicke and Roediger, 2008; Karpicke, 2012; Eisenkraemer et al., 2013). Studies investigating the effects of tests in laboratory settings typically involve the comparison between conditions in which participants take low-stakes tests on previously studied materials (i.e., practice the retrieval of the studied information), to conditions in which participants restudy the previously studied materials. Typically, practicing retrieval yields significantly greater long-term retention of the

studied materials than just restudying them (e.g., Pashler et al., 2007), a phenomenon frequently termed "testing effect" or "test-enhanced learning." This effect has been thought to be a promising strategy to improve learning in classroom settings, and has been recommended as an effective and inexpensive learning strategy to be adopted in schools (Roediger and Pyc, 2012). The question arises, however, whether there is enough evidence for such recommendations. That is, is there a sufficient amount of evidence provided by applied research demonstrating that the testing-effect can promote learning in actual educational contexts? In order to approach this issue, we review here studies that assessed the testing-effect in real educational settings. We focused our analysis on the types of tests these studies used, on the control conditions they adopted, on whether they found corrective feedback beneficial, and most importantly, on whether they were really beneficial for learning in such contexts.

Prior laboratory research have used several types of tests for the practice of retrieval, which can be broadly characterized as free-recall, cued-recall, and recognition tests. In typical free-recall tests, participants are first presented to a series of stimuli (e.g., lists of unrelated words), and are later asked to remember as much of the presented stimuli as they can. In cued recall tests, participants are asked to remember stimuli that were previously associated with specific cues. Thus, in a typical cued-recall paradigm, participants could first see pairs of words in the study phase, as for instance, "house"–"computer." After an interval, they could be asked to remember the word "computer" when the word "house" was presented again ("house"–"_____"). Finally, in recognition tests, participants are asked to discriminate between studied and unstudied materials. For example, participants may study a series of words and in a later moment see the studied words among unstudied words. The participants' task would be to indicate which words were "old" (i.e., studied) and which were "new" (i.e., unstudied). All these tests have been shown to elicit reliable testing effects in laboratory settings (e.g., Carpenter et al., 2006; Roediger and Karpicke, 2006b; Smith and Karpicke, 2014). In educational settings, however, the format of these tests are often different. Free-recall tests often can take the format of exercises wherein students are asked to retrieve previously studied materials without external support; cued-recall can take the format of fill-in-the-gaps or short-answer tests; and recognition tests can take the format of multiple-choice tests.

An important question for both laboratory and classroom research, is whether these different types of tests are differently effective in eliciting testing effects. Contemporary theories of cognitive psychology propose that memory retrieval is based primarily on two processes, namely, a subjective sense of familiarity (i.e., a sense of knowing that a given stimulus was encountered before), and a more detailed recollection of contextual features of prior events. These retrieval processes are termed *familiarity* and *recollection*, respectively (Yonelinas, 2002). Several studies show that performance on free-recall and cued-recall tests are heavily dependent on recollection, whereas performance on recognition tests can be based on familiarity only, or in a combination of familiarity and recollection (Yonelinas and Parks, 2007). Thus, because recollection consists in a more elaborative retrieval of studied information, and involves the reinstatement of episodic/contextual features of the original study event, according to current testing effect theories free- and cued-recall tests are expected to produce greater testing effects than recognition (Carpenter, 2009; Karpicke, 2017). Although several reports are consistent with these prediction (e.g., Stenlund et al., 2016), a recent meta-analysis suggested that this is not always the case (Adesope et al., 2017). The meta-analysis actually shows that recognition tests in the form of multiple-choice tests are at least as effective as free- and cued-recall in reproducing testing-effects. Thus, in the current article we examined whether in educational settings the type of test used for retrieval practice is an important factor for learning. Furthermore, the aforementioned meta-analysis showed that retrieval practice is beneficial for primary, secondary, and post-secondary students, but it did so by collapsing laboratory and classroom studies. Those authors, therefore, did not examine whether retrieval practice is beneficial for students from specific age ranges in actual educational contexts.

Another important question for the implementation of the retrieval practice in classroom environments concerns the type of control condition the retrieval practice is compared to (Kornell et al., 2012). Although the effectiveness of retrieval practice for memory retention is typically assessed by its comparison to a restudy condition (e.g., participants reread the studied materials), the latter is known to be a particularly weak learning strategy (Callender and McDaniel, 2009). Prior laboratory experiments, however, have shown that even when testing is compared to stronger learning strategies, testing remains advantageous (Karpicke and Blunt, 2011). To verify whether such findings are replicable in classroom settings, we describe and discuss the type of control conditions implemented by each reviewed study. Because adopting retrieval practice in classroom settings may entail abandoning other teaching activities, knowledge concerning which strategies are advantageous or disadvantageous in contrast to retrieval practice is particularly important for the well-informed application of retrieval practice in school environments.

Finally, in laboratory settings, corrective feedback has been shown to enhance learning in multiple-choice tests (Butler et al., 2007; Butler and Roediger, 2008) and to enhance learning when participants have low confidence on their responses (Butler et al., 2008). The results of the aforementioned meta-analyses (Adesope et al., 2017), however, suggest that the testing effect was similarly elicited whether feedback was provided or not. Thus, although more laboratory research will be necessary to elucidate the conditions in which feedback may be more or less advantageous for retrieval practice, an important question for educational purposes is whether feedback is beneficial for retrieval in classroom contexts. This question is approached in the current review.

Thus, overall, the central questions of the current review are (1) whether the testing effect is replicable in educational settings; (2) whether there are types of tests that are more (or less) beneficial than others in educational contexts; (3) whether the potential benefits of retrieval practice are more prominent to specific age-ranges in educational settings (4) whether retrieval

practice remains beneficial when compared to "stronger" control conditions in the classroom; (5) whether corrective feedback enhances the benefits of retrieval practice in classroom settings; and (6) based on the answer for the questions above, whether the current applied literature is substantial enough to instill the recommendation of retrieval practice in school environments.

Considering the importance of these questions for educational purposes, it is surprising that so few reviews on this topic have been reported so far. An early meta-analytic review (Bangert-Drowns et al., 1991) showed that testing is advantageous for learning in classroom settings, however, the types of tests used in the reviewed studies and the presence or absence of feedback were not considered by those authors. Three more recent meta-analytic reviews considered the application of the testing effect to educational contexts focusing on the influence of testing on students' achievement (Phelps, 2012; Schwieren et al., 2017), or included studies conducted in classroom settings as a moderator in the analysis (Adesope et al., 2017). Although they were all favorable to the use of retrieval practice in educational contexts, due perhaps to the meta-analytic procedures they adopted, no detailed description of how retrieval practice was applied to educational settings was provided. Specifically, no comparisons were made among the applied testing-effect research regarding issues as type of test, age of students, type of study task, type of course, and so on. Indeed, in these meta-analyses the applied research was considered only as a moderator, and no analysis considering the differences among the selected applied studies were conducted. Two further testing-effect meta-analytic reviews did not approach the testing effect on classroom environments (Rowland, 2014; Pan and Rickard, 2018).

Regarding the narrative reviews conducted so far, the review reported by Roediger and Karpicke (2006b) included a section about the application of the retrieval practice to educational contexts. Because most experimental studies on testing effects were conducted in the last decade, the only articles these authors cite on this section are the aforementioned meta-analysis (Bangert-Drowns et al., 1991) and one empirical study (Leeming, 2002). The remaining narrative reviews do not include analysis focusing on detailed educational applications of the testing effect (Dunlosky et al., 2013; Eisenkraemer et al., 2013).

In general, there are several differences between laboratory and educational environments, as for example, the amount of information students are required to learn, the motivation to learn the studied materials, the way the to-be-learned materials are presented, and perhaps more importantly, the differences in the amount of distraction existing in each of these environments. In spite of these differences, recommendations have been often made in favor of the adoption of retrieval practice in classroom environments, which are based mostly on laboratory studies (Roediger and Pyc, 2012; Dunlosky et al., 2013). Furthermore, no review study have shown so far whether the extant literature concerning the applied testing effect gives sufficient support for the ample use of retrieval practice in educational settings. Thus, in contrast to prior reviews on retrieval practice, the current study has a specific focus on the application of the retrieval practice to educational settings. Because retrieval practice can be applied in many different ways, detailed descriptions of how this application was conducted in each study are crucial for future research, or for future applications of such practices to actual classroom contexts. Thus, in sections dedicated to each type of test, and subdivided according to age (from older to younger students), we first summarize the testing procedures of each study, and then conduct integrative discussions focused on each type of test. Overall conclusions and recommendations for future research are proposed toward the end of the article.

METHOD

In order to select the articles for the present work, we searched the databases *Web of Science, PubMed* and *Google Scholar*, for keywords or titles containing the terms "testing effect," "test-enhanced learning," "retrieval practice," "applied," and "classroom," during the first semester of 2018. To complement the search, further searches were conducted in the reference lists of the selected materials, and in the reference lists of prior reviews on testing effect (e.g., Dunlosky et al., 2013; Adesope et al., 2017). No restriction concerning date of publication was applied. Since the goal was to review studies that were applied to educational contexts only, all articles reporting laboratory studies were excluded. Thus, the inclusion criteria for the selection of studies were (1) articles should present empirical studies; (2) the focus of the experiment should be on the retrieval practice; (3) studies should focus on typically developing individuals; (4) experiments should be applied to actual educational environments in the sense that (a) the to be learned materials were directly related to the content normally exposed and evaluated in particular courses/disciplines, and (b) most phases of the study were conducted in classroom settings, or, in the case of computer-based tests, on platforms frequently used by the studied educational institutions.

RESULTS OF THE LITERATURE REVIEW

After an initial screening based on title and abstract, 67 articles were found from the relevant keywords. These articles were then fully examined according to the inclusion criteria, which resulted in the selection of 23 articles (see **Table 1**), all reporting experiments conducted in actual classrooms (see inclusion criteria above). Below, each selected article is briefly described and discussed, focusing mainly on type of test, control condition, and on whether feedback was provided or not. The subsections are structured according to type of test, moving from "shallower" to "deeper" types of tests (Craik and Tulving, 1975), and then approaching studies that contrasted or combined different types of tests. Within each subsection, we begin by reviewing studies that recruited undergraduate or medical school students as subjects, and then review studies recruiting high school or elementary school children. The rationale for this organization is to approach first the studies that are more similar to the typical laboratory experiments, and examine whether the usual testing-effect is replicated when just the environment is changed (i.e., from the lab to actual educational contexts), and

TABLE 1 | Main procedures used in the reviewed articles.

References	Retrieval practice test	Participants/course	Materials for actual exams?	Individual differences?	Study task	Control condition	Feedback	Was retrieval practice overall beneficial?
Balch, 1998	Multiple-choice	Introductory psychology students	Yes	No	Lectures	Reread and expectancy rating	Yes	Yes
Batsell et al., 2016	Multiple-choice	Psychology students	Yes	No	Assigned readings	No activity	No	Yes
Daniel and Broida, 2004	Multiple-choice	Psychology students	Yes	No	Lectures and assigned readings	No activity	Yes	Yes
Kibble, 2007	Multiple-choice	Medical students	Yes	No	Lectures	No activity	Yes	Yes
McDaniel et al., 2011	Multiple-choice	Science questions applied to 8th students	Yes	No	Assigned readings and lectures	No activity	Yes	Yes
McDaniel et al., 2013	Multiple-choice	Science questions applied to 9th students	Yes	No	Assigned readings and lectures	No activity	Yes	Yes
Roediger et al., 2011	Multiple-choice	Social studies questions applied to 6th graders.	Yes	No	Lectures	Reread or no activity	Yes	Yes
Vojdanoska et al., 2010	Fill-in-the-gaps	First year psychology students	No	No	Power-point presentation	No activity	Yes vs. No	Yes
Jaeger et al., 2015	Fill-in-the-gaps	3rd grade students	No	Yes	Read a 321-words text	Reread	No	Yes
Larsen et al., 2009	Short-answer	Pediatrics and emergency medicine residents	No	No	Lectures	Read review sheets about the lectures	Yes	Yes
Wiklund-Hörnqvist et al., 2014	Short-answer	Undergraduate students enrolled in a Cognitive psychology class	No	No	Assigned readings	Reread	Yes	Yes
Carpenter et al., 2016	Short-answer	Undergraduate introductory biology course students	Yes	Yes	Read the definition of biology terms.	Copy the definitions of the studied biology terms	Yes	Only for high-performance students.
Lyle and Crawford, 2011	Short-answer	Undergraduate psychology students enrolled in a statistic course	Yes	No	Lectures	Lectures	Yes	Yes
Dirkx et al., 2014	Short-answer	Probability calculation for high-school students	No	No	Assigned readings	Reread	Yes	Yes
Carpenter et al., 2009	Short-answer	8th grade history students.	No	No	Discussions, notes, readings and handouts	Reread	Yes	Yes
Lipko-Speed et al., 2014	Short-answer	Tests of science concepts administered to 5th graders	No	No	Read terms and their definitions	Reread or no activity	Yes vs. No	Only when feedback was provided.
Goossens et al., 2016	Short-answer	Vocabulary learning for 2nd, 3rd, 4th, and 6th graders	No	No	Illustrations and conversations with instructor.	Copy the description of words	No	No
Dobson and Linderholm, 2015	Free recall	Anatomy and physiology students.	Yes	No	To read a 600-words text	Reread or taking notes	Yes	Yes
Ramraje and Sable, 2011	Multiple-choice vs. short-answer	Medical students	No	No	Lectures	No activity	No	Yes

(Continued)

TABLE 1 | Continued

References	Retrieval practice test	Participants/course	Materials for actual exams?	Individual differences?	Study task	Control condition	Feedback	Was retrieval practice overall beneficial?
McDermott et al., 2014	Multiple-choice vs. short answer	7th grade science and high school history	Yes	No	Lectures	Reread	Yes	Yes
Leeming, 2002	Short-answer + short-essay	Psychology students	Yes	No	Lectures	Lectures	Yes	Yes
Cranney et al., 2009	Fill-in-the-gaps + Multiple-choice	Psychology students	No	No	Video presentation	Read and highlight a summary of the video or no activity	Yes	Yes
Burdo and O'Dwyer, 2015	Short-answer + Multiple-choice	Psychology students	Yes	No	Lectures	Concept mapping or no activity	Yes	No

subsequently verify whether the effects are also replicated for younger participants.

Thus, below we begin by discussing the current evidence (if any) in favor of the application of multiple-choice tests to educational settings.

Do Multiple-Choice Questions Increase Learning in the Classroom?

Multiple-choice questions are frequently used to evaluate learning, since they are easy to administer and easy to grade. The question then is whether multiple-choice tests can be a useful strategy to enhance learning in the classroom. We found four studies examining this question in undergraduate and medical students, and three studies examining this question in elementary school children.

In the study conducted by Balch (1998), undergraduate students in an introductory psychology class were divided in two groups, the "practice-exam" group and the "review-exam" group. Participants in the practice-exam group received a multiple-choice test with questions similar in format and content, but not identical, to the questions they would later encounter in a final exam. Responses to such questions were followed by corrective feedback. Participants in the review-exam group were asked to read the same questions, along with their answers, and to rate them in a 4-point scale regarding their expectancy of finding these questions again in the final test. In the final test, a week later, students from the practice-exam group exhibited superior performance than participants from the review-exam group.

Batsell et al. (2016) assigned undergraduate students enrolled in an introductory psychology course to a control group, which received daily readings; and to a quiz group, which received daily readings and completed daily multiple-choice quizzes on the readings. Students from both groups performed three exams containing multiple-choice questions throughout the semester. The questions of the exams were divided in three types: identical questions (identical to the quizzed questions), new questions, and similar questions. The similar questions covered topics that had been quizzed, but were substantially different from the questions of the quizzes applied before. The use of daily quizzes improved performance on identical, similar, and new questions compared to the no-activity ("control") condition. Notably, feedback was not given in this study, and the testing effect was assessed only for materials studied in the textbook (i.e., not studied in the classroom).

In the study reported by Daniel and Broida (2004), undergraduate students enrolled in psychology classes were assigned to one of three conditions: no quiz, in-class quiz, and web-based quiz. In the in-class quiz condition, students responded to quizzes (i.e., 10 chapter-based questions) during the first 15 min of each class. In the web-based quiz condition, the same quizzes were available online for the students during the 24 h preceding each class (for 15 min once they get started). In both conditions, students received immediate feedback of their performance on each quiz. In the no quiz condition, students did not complete any kind of quiz, nor did they have access to them. Performance was assessed in four multiple-choice and/or

short-answer tests throughout the semester. As expected, the in-class quiz group performed significantly better than the no-quiz group. However, the performance of the web-based quiz group did not differ significantly from the no-quiz group. The authors later discovered that students were resorting to various strategies to cheat while taking the online quiz. Thus, to avoid this, some changes were made in the online quiz. Specifically, the 10 questions of each quiz were randomly selected from a pool of 100 questions, and the time to complete each quiz was reduced from 15 to 7 min. After those changes, both experimental groups (i.e., web-based quiz and in-class quiz) performed similarly and significantly better than the no quiz group in two further examinations.

Kibble (2007) also examined whether online quizzes enhance learning, but also examined whether the use of quizzes could be increased through a reward system. Quizzes were available to students from five different classes of physiology (approximately 350 students per class) to get them prepared for their usual middle- and end-of-semester exams. Two quizzes were available before each exam, and corrective feedback was given after the second quiz. For each of the five classes, the quizzes were offered within different reward models. In model 1, no incentive was given for the completion of the quizzes; in model 2, 0.5% of the grade was given for those who accessed the quizzes; in model 3, 1% of the grade was given for those who responded correctly to 30% of the quizzes or more; in model 4, 1% of the grade was offered according to the scores obtained in the best of two attempts to complete the quizzes; and in model 5, 2% of the grade was given using the same criteria of model 4. In the class of model 1, as no incentive was offered for the completion of the quizzes, only 52% of the students performed them. Those who completed at least one quiz in this group, performed significantly better than those who did not perform any quiz (i.e., a "no activity" control condition). In the other classes, incentives increased dramatically the access to quizzes, resulting in a very small number of students who did not take any quiz. Interestingly, greater incentives led some students to take quizzes incorrectly, using previously written notes or the help of other people. These students, who achieved a score close to 100% in their first attempt, did not achieve a grade as good as the students who used the two attempts to complete the quiz.

The four studies investigating whether multiple-choice tests are beneficial for medical school students and for undergraduate students show positive results. Below, the three studies assessing multiple-choice tests in children are reviewed.

In the study reported by McDaniel et al. (2011) there were three experiments in which 8th grade students performed three consecutive multiple choice "quizzes" on science contents, followed by corrective feedback. The first quiz was applied before class (students had been instructed to read a text about the lesson at home), the second quiz was applied immediately after class, and the third quiz was applied 24 h before each unit exam (20 days after the first quiz, on average). Overall, participants who were quizzed showed increased performance in the unit, end-of-semester, and end-of-year exams in comparison to participants who were not quizzed. Interestingly, quizzes administered 24 h before the unit exams increased performance not only in the unit exams, but in the end-of-semester exams as well. This effect persevered even when earlier quizzes were absent, suggesting that conducting multiple-choice tests before exams can be an effective tool to improve both exam performance and long-term retention. The findings of McDaniel et al. (2011) were replicated in a follow-up study (McDaniel et al., 2013) in which 9th graders were recruited and the format of the questions were different on each of the three quizzes (i.e., some questions focused on applications of the concepts and others on the definitions of the concepts). As in McDaniel et al. (2011), students received feedback after each quiz, and retrieval practice was compared to "no activity." Positive effects of quizzing were found for questions focusing on both the application and the definition of the studied concepts.

Roediger et al. (2011) conducted three experiments in 6th grade social studies classes. In the first experiment, a series of multiple-choice quizzes were administered: (a) pre-test quizzes, administered immediately before class; (b) post-test quizzes, administered immediately after class; (c) review tests, administered a few days after class; (d) chapter exams, administered about 2 days after each review test; and (e) a late exam, which was administered at the end of the semester (about 1 or 2 months after study). The quizzes (pre-test, post-test, and review test) covered all the taught materials and consisted of 4-alternative multiple-choice tests, followed by corrective feedback. The chapter exams consisted of an initial free recall test covering all contents of the chapter, followed by multiple-choice questions about quizzed and unquizzed items. Students exhibited greater performance at the chapter exams for quizzed than for unquizzed materials, and such enhanced performance was replicated in the later exams. In the second experiment, a control group that just read the quiz questions was included, and the enhanced retention of quizzed materials found in experiment 1 was replicated in the chapter exam, and partially replicated in the late (end-of-semester) exam. Finally, in the third experiment, students were encouraged to use a website to test their learning. The website could be accessed from the students' homes, and offered different kinds of tests, including games requiring responses to relevant questions. This experiment differed from experiments 1 and 2 by keeping only pre-test quizzes (i.e., post-test and review tests were removed). In total, eight pre-test quizzes were given throughout the year. In the chapter exams (multiple-choice and short-answer) and in the end-of-semester exam (multiple-choice, administered 1–3 months after the target materials were initially studied), the effect of test persisted for pre-tested items, despite of whether participants used the website or not. Overall, this study shows that taking multiple-choice tests can benefit later performance in multiple-choice and short-essay tests, suggesting that such practice is flexible in the sense that practicing retrieval in one type of test can enhance performance in a different type of test.

In summary, we found seven studies examining whether multiple-choice tests were beneficial for learning in classroom settings either for young adults (4 studies) or children (3 studies). In all seven studies, multiple-choice tests were beneficial for learning. From these studies, only one did not include feedback (Batsell et al., 2016), and only two studies compared retrieval practice to a reread condition (Balch, 1998; Roediger et al.,

2011), while the remaining five studies compared retrieval practice to "no-activity." Interestingly, in all seven studies the retrieval practice questions differed from the questions used to probe learning. Some differed in terms of the final test adopted (regular exams, short-essays, short-answers), and others only in the manner the questions were exposed (e.g., application vs. definition of concepts). When these differences were directly approached, testing effects were equally strong for both different and identical questions (e.g., McDaniel et al., 2013; Batsell et al., 2016).

In spite of the overall favorable results of multiple-choice tests for students of various ages, because multiple-choice was compared only to "no-activity" or to reread, it is still unclear whether it is as beneficial for learning as alternative learning strategies (e.g., concept mapping). Furthermore, in the only study without feedback, multiple-choice was compared to "no-activity." Thus, the question remains whether multiple-choice tests alone (i.e., without feedback) are beneficial when compared to an actual control condition. This is an important issue to be addressed by future research, and will be further approached in the discussion section.

Does Asking Students to Fill in Gaps Increase Learning in the Classroom?

The studies reviewed above suggest that multiple-choice tests are beneficial to learning in classroom settings. Does a slightly more "difficult" type of test elicit similar results? In the studies reviewed below, students were required to fill-in-the-gaps of texts. Such test is analogous to cued-recall tests, a type of test known for eliciting strong testing effects in laboratory settings (Kang et al., 2007). We found only two studies examining this type of test in educational contexts.

The first was reported by Vojdanoska et al. (2010), involved undergraduate students, and was divided in three phases. In the first phase participants attended to a 10 min long Power-point presentation about adult development. The presentation included short videos and contained at least 24 relevant items of information. In the second phase, which was immediately after the first, participants responded in groups or individually to a test containing 16 fill-in-the-blanks questions (out of the 24 relevant items). Importantly, half of these questions were followed by feedback. In the third phase, which was held a week later, participants responded to a final test, which included the 16 questions presented in the second phase and the remaining 8 untested questions. A further group of participants (control group) skipped the first and second phases, and performed just the final test. The data showed that students who were subjected to the first and second phases showed greater performance than the control group in the final test. More importantly, participants showed greater performance for questions that had been tested in comparison to untested questions, and greater performance for questions followed by feedback than for questions not followed by feedback. Conducting tests individually or in groups resulted in indistinguishable performance at the final test.

In the second study investigating fill-in-the-gap tests in educational contexts (Jaeger et al., 2015), 3rd grade children read a 321-word encyclopedic text about the Sun, which contained 20 key terms. After reading the text twice, students either read the text twice again, or received the text with the 20 key terms missing, and had to write them down in the appropriate gap (e.g., The superficial layer of the sun is called _____ [*photosphere*]). Importantly, corrective feedback was not provided in this study. Seven days later, in the final multiple-choice test, students in the "fill-in-the-gaps" condition had a very superior performance. Also, measures of individual differences in IQ and reading skills conducted in this study suggest that practicing retrieval can benefit any children within the normal range of these abilities.

In summary, we found only two studies investigating the impact of fill-in-the-gaps tests on the retention of educational materials. One of these studies recruited undergraduate students and the other 3rd grade children. Both showed positive testing effects after a 1 week interval. Beyond the differences between these studies in terms of the age of the participants, they also differed in terms of the learning materials (i.e., Power-point presentation vs. an encyclopedic text), in terms of the restudy condition adopted, and in terms of whether feedback was present or absent. Regarding the control condition, Jaeger et al. (2015) compared retrieval practice to rereading, while Vojdanoska et al. (2010) compared retrieval practice to "no-activity." Feedback, on the other hand, was absent in the study reported by Jaeger et al., and was an independent variable in the study reported by Vojdanoska et al. The results of the latter showed that the benefits of testing were enhanced when feedback was provided. Thus, although the number of studies assessing the application of fill-in-the-gap tests to actual classrooms remains small, this simple retrieval practice seems promising, especially taking into account that it is easy to apply, it is not time consuming, and it is relatively easy to correct and to provide feedback.

Do Short-Answer Tests Increase Learning in Classroom Settings?

As fill-in-the-gaps tests, short-answer tests are analogous to cued-recall tests, which is often used to study memory in laboratory settings. Such tests are thought to elicit stronger testing-effects relative to multiple-choice tests (Carpenter, 2009; Karpicke, 2017; although see Adesope et al., 2017). Thus, we expected that studies using short-answer tests during retrieval practice would show stronger testing-effects in classroom settings. We found eight studies examining the potential benefits of using short-answer tests in classroom settings. Four of the eight studies recruited young adults as participants, one recruited adolescents (high-school students), and three recruited elementary school children.

Among the studies with young adults, one study recruited medical residents (pediatrics and emergency) and three recruited undergraduate students enrolled in biology and psychology classes. In the study in which medical students were recruited (Larsen et al., 2009), participants attended to lectures about myasthenia gravis and epilepsy. Half of the participants were tested on epilepsy, and restudied (i.e., read review-sheets about the lectures) myasthenia gravis, while the other half was tested on myasthenia gravis and restudied epilepsy. Test and restudy sessions were held both immediately after the lectures, and

after a 2 weeks interval, and they were followed by corrective feedback (i.e., a sheet with the correct answers). In a final test, administered 6 months after the second restudy/test session, materials tested with short-answer questions were significantly more remembered than restudied materials.

The study reported by Wiklund-Hörnqvist et al. (2014) verified whether short-answer tests are more beneficial than restudy for the retention of cognitive psychology concepts. After learning 57 cognitive psychology concepts through assigned readings, a lecture, and a further short reading about the concepts, undergraduate students (enrolled in a cognitive psychology class) either restudied each concept and its respective definition (i.e., reread it), or remembered and then typed the name of the each concept after reading its definition. Both restudy and retrieval practice were repeated six times for each concept and were performed individually in a computer. Corrective feedback (i.e., the actual correct answer) was provided after each short-answer response. Retrieval practice showed greater retention than restudy in short-answer tests administered 5 min, 18 days, and 5 weeks after the retrieval and restudy sessions were finished.

In the study reported Carpenter et al. (2016), introductory biology undergraduates performed terms' definition exercises that were followed by the recall or the copy of the studied terms' definitions (e.g., Polar body: a cell produced by asymmetric cell divisions during meiosis). After their responses, a corrective feedback was given. Five days later, all students completed an unexpected quiz assessing the information learned from the exercises. Their results showed that only high-performing students (as assessed by prior course exams) were benefited by recalling the definitions of the terms, while low- and middle-performing students were more benefited by copying the definitions than recalling them.

Lyle and Crawford (2011) assessed the potential effects of short-answer questions provided in the end of each class of a statistics course. The authors of the study taught this course to psychology students in two consecutive years. Tests were administered only at the second year, and performance at the first year served just for comparison. The short-answer questions were projected on a whiteboard during the last 10 to 15 min of each class. Students were required to write down and submit their responses before leaving class. At the beginning of each class, the correct answers (i.e., feedback) for the questions from the previous class were projected on the whiteboard (and posted on the course website), and participants were free to review their responses and make further questions before new materials began to be taught. In order to motivate students to perform the tests, they counted for 8% of the course grade. Performance of students in the comparison group (without tests in the end of the classes) and students in the test group were assessed by four multiple-choice tests distributed throughout the semester. The exam questions did not use the same wording as the test questions, but their main contents were the same. In three out of the four examinations of the course, participants who did take tests performed significantly better than participants who did not. In addition, it was also found that the students' scores in the end-of-the-class tests were positively correlated with the grades obtained in the exams.

The four studies assessing the use of short-answer tests as a learning tool for undergraduates and medical school students showed promising results. They all showed positive results, except for Carpenter et al. (2016), who found that only students who exhibited high performance in prior exams were benefited by performing short-answer tests. This is an important finding, which will be further approached in the Discussion section. Below, we turn to studies examining the application of the testing effect in high school (1 study) and elementary school (3 studies).

In the study with high-school students (age-range = 15 to 16 years), Dirkx et al. (2014) examined whether retrieval practice benefits the learning of principles and procedures of probability. The students were assigned to a restudy or a retrieval condition. Both groups first read an 899-words text about probability calculations. The participants assigned to the restudy condition reread the same text three times more, performing in total four study sessions (SSSS). The participants assigned to the retrieval practice condition, performed tests after the initial reading, read the text again, and performed the same tests a second time (STST). Note that corrective feedback in this case consisted in the second study phase. The tests were 10 short-answer questions in which participants either applied to a novel situation a principle/procedure of probability calculation read in the original text, or remembered factual information read in the text. Thus, participants assigned to the retrieval practice condition performed both types of test. A posttest containing factual knowledge and principle-application questions (short-answer) were administered again 1 week after the learning phase was finished. Performances on both types of test (factual or application of principles) were equivalent; nonetheless, both tests were significantly more beneficial than restudy.

In the study conducted by Carpenter et al. (2009), 8th grade students responded to short-answer questions about history contents. The questions comprised materials exposed to students during class, in discussions, notes, reading assignments, and handouts. The tests (and restudy) were administered 1 week ("immediate review" group) or 16 weeks ("late review" group) after the first exposure of the students to the to-be-learned materials. The tests were in the form of simple history questions (Who assassinated President Abraham Lincoln?) to which participants should provide a short response (John Wilkes Booth) followed by corrective feedback (i.e., the correct response). Restudy was identical, except that the correct answer was shown all the time and participants just reread it. A final test administered 9 months after the students had completed the review, showed increased long-term retention for tested items relative to restudied or studied-only items, an increase that was even greater for the 16 weeks relative to the 1 week group.

The study reported by Lipko-Speed et al. (2014) involved 5th graders and focused on the science contents that are typically studied in this grade, namely, light and sound (experiment 1), and geography (experiment 2). This study comprised four sessions. The first and second sessions were administered with an interval of 48 h between them (on Monday and Wednesday of the same week). The third and fourth sessions were administered 1 week later, also with a 2 day interval between them (on

Monday and Wednesday of the following week). In the first session, all participants studied 20 key terms and their definitions (e.g., What is sound? *Form of energy that you can hear that travels through matter as waves*). Five of the 20 key terms were then assigned as controls, and not presented again until the fourth session (final test). During the second and third sessions, the remaining 15 key terms were assigned to three conditions (five to each condition). In the "test" condition, participants were required to type the definitions of each key term (What is sound? _____); in the "test-plus-feedback" condition, participants received feedback after typing the definition of each key term, and in the "study" condition, each definition was presented again, and participants reread them in a self-paced manner. Finally, in the fourth session (final test), students were asked to type the concept of all 20 key terms. Items assigned to the test-plus-feedback condition were better remembered in the final test than items that were just tested or restudied. Surprisingly, however, the contrast between the recall of tested items (without feedback) vs. restudied items in the final test was not significant, suggesting that feedback was essential to increase learning.

Finally, the study reported by Goossens et al. (2016) investigated whether remembering the description of words was more beneficial for vocabulary learning of 2nd, 3rd, 4th, and 6th graders than copying the descriptions of words. The experiment started with the children receiving the definitions of the words through illustrations and conversations with the experimenter. After this, the children performed several textbook exercises with the learned words, including exercises involving copying (restudy) and remembering (retrieval practice) each word description. The exercises were distributed over 2 or 4 weeks, and the effectiveness of retrieval practice was assessed by short-answer tests administered in the end of each week, and by a final multiple-choice test administered from 1 to 11 weeks after the exercises were finished. No corrective feedback was provided. In both short-answer and multiple-choice tests, retrieval was no more beneficial than restudy. Actually, for the 3rd grade children restudy was even more beneficial than retrieval in both tests.

Thus, from the four studies with high school and elementary school children, two were not very favorable to the use of retrieval practice in class. More specifically, in one study with 5th graders the testing-effect was beneficial only when followed by feedback (Lipko-Speed et al., 2014), and in another it was not reliably beneficial for children from different grades (i.e., 2nd, 3rd, 4th, and 6th graders; Goossens et al., 2016). This raises the question of whether short-answer tests are indeed a beneficial learning strategy for elementary school children. Note that the remaining two studies showing positive testing effects for children included actually 8th graders (Carpenter et al., 2009) and high school adolescents (Dirkx et al., 2014). Thus, more research will be necessary before recommendations of short-answer tests can be made for children below the 8th grade.

Considering all short-answer studies, the differences between studies that did elicit testing-effects relative to studies that did not seem to be related to the type of restudy condition. Note that rereading (Carpenter et al., 2009; Dirkx et al., 2014; Wiklund-Hörnqvist et al., 2014), reading summaries of lectures (Larsen et al., 2009), or attending to the final minutes of a lecture (Lyle and Crawford, 2011) were among the control conditions of the studies showing positive testing-effects. When tests were compared to activities involving copying written materials, retrieval practice was not advantageous (Goossens et al., 2016), except for high-performance students (Carpenter et al., 2016). These results highlight the importance of the control condition to determine whether retrieval practice is successful, and brings into question the capability of retrieval practice to yield greater retention than alternative learning conditions (i.e., learning conditions other than restudy or "no-activity"). This possibility is further approached in the discussion section.

Finally, feedback was provided in all studies showing positive testing-effects. Conversely, the only study contrasting the presence vs. the absence of feedback showed that the testing effect occurred only when feedback was present (Lipko-Speed et al., 2014), and the only study not adopting feedback showed no testing-effects (Goossens et al., 2016). Although the reason for the absence of reliable testing effects here may be related to the age of the participants (see above), these findings suggest that feedback may play an important role when short-answer tests are used for retrieval practice, although as discussed later, more research will be necessary to assess this possibility.

Does Free-Recall Benefit Learning in Classroom Settings?

In free-recall tests, individuals are asked to produce (i.e., recall) previously studied information without the help of cues. Such type of test has been shown to produce strong testing effects in cognitive psychology laboratories (e.g., Roediger and Karpicke, 2006b). To assess the effectiveness of such type of test in the classroom is rather challenging, however. The materials used in class are typically presented in a less controlled manner than it is presented in laboratory settings. That is, while in laboratory settings participants might encode lists of words, often controlled for concreteness, frequency, number of letters, which are presented individually and for a predetermined amount of time; in the classroom, materials are often presented in lectures, texts, and discussions, which are conducted in groups including several students.

Perhaps because of these difficulties, only one study to date investigated whether conducting free-recall tests could be an effective strategy to improve learning in classroom settings. The study was reported by Dobson and Linderholm (2015), and involved two phases. In phase 1, students of an anatomy and physiology course studied one out of three passages describing structures and concepts of (1) cardiac electrophysiology, (2) ventilation, and (3) endocrinology. Each passage had a little more than 600 words, and the study of each passage was conducted in different conditions. That is, each student (a) read one of the passages three times in a row (R-R-R), (b) read another of the passages and then reread it while taking notes (R-R+N), and (c) read the remaining passage, completed a free-recall task, and then read it again (R-T-R). The free-recall task consisted of writing down in a blank sheet as many concepts and definitions from the passage as possible. Immediately after this, and once again 1 week later, participants performed multiple-choice tests on the studied contents. In the first test (immediate), no significant difference

in performance between the R-T-R and R-R+N conditions was found, although performances on both conditions were better than performance in the R-R-R condition. In the delayed test (1 week later), however, performance in the R-T-R condition was greater than performances in the other two conditions. In the second phase of the study, the results of the first phase were presented to the participants, and they were encouraged to use the superior R-T-R strategy to get prepared for the remaining course exams. The use of such strategy significantly increased performance on the course exams, suggesting that when students adopt self-testing in their preparation for exams, learning can be significantly improved.

Even though a large number of laboratory studies posits free recall as an excellent practice to enhance long-term retention (Dunlosky et al., 2013), this is the only study examining free-recall in actual educational settings we found. It included two control conditions, namely, a restudy only condition (R-R-R) and a study-plus-taking notes (R-R+N) condition, and the test condition was more beneficial than both when memory for the studied materials was assessed 1 week later (see Roediger and Karpicke, 2006b, for similar results in laboratory settings). Because a retrieval condition without feedback was not included (e.g., R-T-T), it is not possible to infer whether feedback was actually necessary for the success of the free-recall retrieval practice (i.e., would this effect persist without feedback?). Thus, future work should examine the role of feedback for this type of test, as well as examine whether the benefit of retrieval practice over the other learning strategies persist after longer intervals.

Contrasting Different Types of Tests

The two studies reviewed below examined separately the effects of different types of test on learning. More specifically, they compared the testing effects produced by short-answer vs. multiple-choice tests. Such comparison can be highly informative since it can yield a clearer notion of which of these tests (if any) is more beneficial to learning in classroom settings.

In the study reported by Ramraje and Sable (2011), medical school students enrolled in Pathology classes were assigned to one of three treatments: (1) multiple-choice tests administered in the end of the class, (2) short-answer tests administered in the end of the class (3), absence of tests in the end of the class. After 3 weeks, all students completed an unexpected test comprising multiple-choice and short-answer questions. As expected, both multiple-choice and short-answer tests resulted in greater performance relative to no test in the later tests. Surprisingly, however, in both final tests participants who performed multiple-choice tests showed greater performance than participants who performed short-answer tests.

The study reported by McDermott et al. (2014) comprised a series of experiments involving 7th-grade science and high school history contents. In all experiments, students took intermittent quizzes (short-answer or multiple-choice, both with corrective feedback), and performance on unit-exams and end-of-semester exams were analyzed. The quizzes were administered immediately before the materials were taught (prelesson quiz), immediately after the materials were taught (postlesson quiz), and 1 day before the unit exam (review quiz), which took place a few days after the postlesson test. Both multiple-choice and short-answer questions were projected on a screen in front of the classroom, and the research assistant read the questions aloud to the students. The students responded within a limited amount of time by using clickers or writing down their responses in a paper sheet. The restudy condition consisted in projecting the question with its most appropriate answer. The results were clearly favorable to practicing retrieval relative to restudying the materials. More importantly, short-answer and multiple-choice questions were equally effective in enhancing the retention of the studied materials.

Overall, from the studies examining multiple-choice vs. short-answer tests, one did not include feedback and compared retrieval practice to "no-activity" (Ramraje and Sable, 2011), while the other did included feedback and compared retrieval practice to a reread condition (McDermott et al., 2014). Even though both studies were favorable to the use of multiple-choice and short-answer questions relative to restudy or no-activity, multiple-choice tests showed some advantage over short-answer in one of the studies (Ramraje and Sable) and was equivalent to short-answer in the other (McDermott et al.). Taking into account the positive results encountered in the studies using exclusively multiple-choice questions (see section Do Multiple-Choice Questions Increase Learning in the Classroom?), and the application and grading advantages of this type of test, such strategy arises as a promising approach to be used in classroom settings.

Combining Different Types of Tests

The studies below assessed whether the concomitant use of different types of tests in the retrieval practice condition (e.g., short-answer + short-essay) is beneficial for learning in classroom settings. Even though such combinations may be useful in classroom contexts, they may preclude a clear notion regarding whether a specific type of test is more or less beneficial than other types of test. In spite of this experimental caveat, we discuss below the three studies we found that adopted such approach. Noticeably, all the three studies recruited undergraduate students in psychology and physiology classes.

Leeming (2002) investigated whether short-answer and short-essay tests were beneficial for learning in two regular classes of an introductory psychology course (i.e., Introductory Psychology, and Learning and Memory). While half of the students performed only the usual three examinations during the semester, the remaining students performed daily tests during the whole course. The daily tests lasted from 10 to 15 min and consisted of two short-essay questions taken from the book adopted in the course, and five short-answer questions based on the content of the texts and the content of the lessons. Participants in the control condition had regular classes during the retrieval practice period (i.e., the last 10 or 15 min of each class). Immediately after each daily test, the researcher commented and corrected the inappropriate responses (i.e., provided corrective feedback). All participants took a final test about 6 weeks after the classes were finished. The final test contained short-essays, multiple-choice, and fill-in-the-gaps questions about the issues covered by the daily tests. The results

showed that the students who performed daily tests had better performances on the final test than the students who did not take them.

Cranney et al. (2009), assessed the benefits of "fill-in-the-gaps" and multiple-choice tests in two experiments. The first experiment was divided in three phases: (1) initial learning, (2) review and (3) final memory test. In the initial learning phase (1), all students watched a video about Psychobiology, which consisted in an introduction on brain signaling and brain structures. In the review phase (2), students were assigned to one of the following groups: (a) groups of 4 to 5 people performed a quiz about the video collaboratively, (b) students completed the same quiz individually, (c) students read a summary of the video and highlighted the most important parts and had 2 min to ask any questions, and (d) students did not re-engage with the video information (i.e., "no-activity"). In the conditions including responding to quizzes (i.e., groups a and b), students had to answer multiple-choice and fill-in-the-gaps questions about the film (e.g., the hippocampus is part of the _____ system [*limbic*]), and all their answers were followed by corrective feedback. In the final memory test (3), 1 week after the review phase, participants who performed quizzes in groups showed greater performance than participants who performed quizzes individually (and from those who performed no quiz). In experiment 2, the group quiz condition was excluded, and participants in the individual quiz condition showed greater performance than participants in the restudy or "no-activity" conditions.

Burdo and O'Dwyer (2015) assessed the potential benefits of short-answer and multiple-choice tests for learning. In their study, undergraduate students enrolled in a physiology class were assigned to three groups. Group 1, attended to 12 encounters during the semester to review the studied materials using the concept mapping method (Novak, 2010; see also Karpicke and Blunt, 2011). Group 2 attended to 12 encounters in which the retrieval of the studied materials was practiced. To perform the retrieval practice, students were divided into small groups and had to elaborate short-answer and multiple-choice questions using the available written materials as much as they felt necessary. They then exchanged the questions with each other so that other members of the group could answer the questions (without access to written materials). The questions then returned to their authors to be corrected, and then correct and incorrect answers were discussed among the group members (i.e., corrective feedback). Group 3 (control group) did not attend to such extra-class meetings as the other two groups did. Learning in each of these conditions was assessed through the course's final grade and five tests placed during the semester (4 unit tests covering materials studied in specific units and 1 cumulative final test). The analyses showed that in the units' tests and in the final grade, students in the retrieval practice group (i.e., group 2) showed no significant higher performance than the other two groups, except in the second unit test, and only when it was considered separately. Furthermore, even though no significant differences were found in the final exam, the control group had a numerically greater performance than the other two groups in this exam.

In sum, three studies examining the use of two types of test in combination were found. In all of them, undergraduate students were recruited as subjects, and corrective feedback was provided. Benefits of testing were yielded when retrieval practice was compared to attending to the remaining minutes of a lecture (Leeming, 2002) or compared to read and highlight a summary of a video presentation (Cranney et al., 2009). When retrieval practice was compared to concept mapping, however, no reliable effects of test were yielded (Burdo and O'Dwyer, 2015). Thus, reliable effects of retrieval practice were found only when relatively "weak" control conditions were adopted. Furthermore, these results are uninformative regarding the importance of feedback, because a condition "without feedback" was not adopted by any of them.

DISCUSSION

The reviewed studies are overall encouraging about the use of retrieval practice to enhance learning in classroom settings. Reliable testing-effects were yielded for all experimental conditions in 19 of the 23 reviewed studies (see **Table 1**). Thus, the response to the first and more general question of the current study, which was whether the testing effect is reproducible in educational settings, is overall positive. Below we discuss the remaining questions raised in the introduction, namely, whether there are types of tests that are particularly beneficial to enhance learning, whether retrieval practice is advantageous for different age ranges in school environments, whether retrieval practice remains beneficial when compared to different types of control conditions, and whether feedback is useful to enhance testing effects in classroom settings. After discussing these questions, we turn to discuss whether there is enough evidence for the recommendation of retrieval practice for actual educational contexts, and discuss further issues observed in the reviewed studies that are potentially important for future research and for future applications of retrieval practice to classroom settings.

Does the Type of Test Matter?

Among the reviewed studies, 8 studies used short-answer tests, 7 used multiple-choice tests, 3 used two types of test combined (unfortunately, without comparing them), 2 contrasted short-answer to multiple-choice tests, 2 used fill-in-the-gaps, and 1 used free-recall. Positive testing effects were yielded for all types of tests. More importantly, however, only studies using short-answer tests (Lipko-Speed et al., 2014; Carpenter et al., 2016; Goossens et al., 2016) or short-answer along with multiple-choice (Burdo and O'Dwyer, 2015) showed absence of testing effects (see **Table 1**). Furthermore, from the 2 studies directly comparing short-answer to multiple-choice, one found no difference between these tests (McDermott et al., 2014), while the other found that multiple-choice tests yielded actually greater retention than short-answer (Ramraje and Sable, 2011).

The absence of testing-effects for short-answer tests found in some studies (e.g., Goossens et al., 2016), or the disadvantage of such test relative to multiple-choice found in one study (Ramraje and Sable, 2011), are inconsistent with the predictions of current theories of retrieval practice (Carpenter, 2009;

Karpicke, 2017), and inconsistent with several laboratory findings (Carpenter and DeLosh, 2006; Stenlund et al., 2016). Notably, such inconsistencies may be caused by differences between educational and laboratory contexts. Perhaps because in educational contexts students are more exposed to distractions, such distractions preclude the more controlled memory search needed for cued-recall tests to enhance memory retention (Karpicke, 2017). Recent findings, however, suggest that cued-recall tests under divided attention conditions still elicits reliable testing effects (Mulligan and Picklesimer, 2017), a finding that is at odds with this possibility. Alternatively, the reason for absence of testing effects in a subset of the reviewed short-answer studies may be related to further characteristics of the studies, as for example, the age of the participants or the type of control condition adopted by the researchers.

Is Retrieval Practice Beneficial for All Ages?

Fourteen of the 23 reviewed studies involved undergraduate or medical school students, and retrieval practice was beneficial in most cases, except in Carpenter et al. (2016) wherein it was beneficial for high performers only; and in Burdo and O'Dwyer (2015) wherein it was not beneficial whatsoever. In both these studies showing absence of benefits, short-answer tests were used during retrieval practice. Interestingly, when multiple-choice and fill-in-the-gaps tests were administered instead of short-answer tests, individuals from all age ranges were consistently benefited. Considering the 8 studies involving children and the one study involving high school students, short-answer tests were not beneficial for children from the 6th grade and below (Lipko-Speed et al., 2014; Goossens et al., 2016), when feedback was not provided. A reason for this resides perhaps in the difficulties that short-answer tests present for children. As mentioned above, short-answer tests (i.e., cued recall) involve a controlled memory search, which can be impaired by distractions, especially in children, since their cognitive control abilities are still developing (Davidson et al., 2006). Future research should verify whether such performance deficiency in controlled memory tests is further replicated in children, elucidating whether such tests can be beneficial for children when feedback is not provided.

Retrieval Enhances Learning in Comparison to What Control Conditions?

As can be seen in **Table 1**, among the reviewed studies, the most common control conditions were "no activity" (11 studies) and rereading (8 studies). Further than that, 2 studies compared retrieval practice to attending to lectures (i.e., the final minutes of lectures), 2 to copying written materials, 1 to taking-notes, 1 to reading review sheets of attended lectures, 1 to read and highlight the summary of a video, and 1 to using concept mapping.

Interestingly, when retrieval practice was compared to rereading or "no-activity" only one study involving 5th graders failed to elicit testing-effects, and only when feedback was absent (Lipko-Speed et al., 2014). Thus, testing-effects seem to be easily yielded in classroom settings when retrieval practice is compared to "no-activity" or to rereading. Notably, however, the comparison between retrieval practice and rereading is somewhat problematic, as rereading has been shown to be a considerably weak learning strategy (e.g., Callender and McDaniel, 2009). The comparison of tests with "no activity" has an even greater limitation, that is, participants are more exposed to studied materials in the retrieval practice than in the control condition (Kornell et al., 2012). Thus, the comparisons between retrieval vs. rereading and retrieval vs. "no activity" conducted in the reviewed studies show that practicing retrieval is more advantageous than using particularly weak learning strategies, and more advantageous than using no strategy whatsoever.

Among the 4 studies showing absence of testing effects, 2 compared retrieval to copying written materials (Carpenter et al., 2016; Goossens et al., 2016), 1 compared retrieval to concept mapping and "no-activity" (Burdo and O'Dwyer, 2015), and 1 compared retrieval to rereading or "no-activity" (Lipko-Speed et al., 2014). Therefore, 3 out of the 4 studies failing to show testing-effects used learning strategies that are evidently stronger than "no-activity," and possibly stronger than repeated reading. A potential conclusion from these findings is that in classroom contexts retrieval practice is as beneficial for learning as traditional classroom activities.

Alternatively, however, the failure of these studies in revealing testing effects may have been caused by limitations inherent to their designs. For instance, in the study reported by Burdo and O'Dwyer (2015), reliable testing effects were absent when retrieval was compared to concept mapping, but also when retrieval was compared to "no-activity," suggesting that the short-answer tests conducted by those authors were perhaps particularly inefficient in eliciting testing effects. In the study reported by Goossens et al. (2016), on the other hand, retrieval practice and restudy were interspersed with additional classroom exercises, precluding the attribution of the absence of testing effects exclusively to the conditions of interest. Thus, future research will be necessary to verify whether the use of retrieval remains advantageous when it is compared to activities that involve deeper levels of encoding (Craik and Tulving, 1975) than activities such as rereading or "no-activity."

In sum, testing effects were clearly yielded in classroom settings when retrieval practice was compared to relatively "weak" learning strategies (e.g., rereading, highlighting, attending to a few minutes of lectures), or when compared to "no-activity." Testing effects failed to be yielded when retrieval was compared to concept mapping (Burdo and O'Dwyer, 2015), even though such comparison elicited robust testing effects in laboratory settings (Karpicke and Blunt, 2011); and failed to be yielded when retrieval was compared with activities involving copying written materials (Carpenter et al., 2016; Goossens et al., 2016). Future research should verify whether these failures are consistent, even when other types of tests are used (e.g., free-recall, multiple-choice).

Is Feedback Useful to Promote Testing-Effects in Classroom Settings?

Feedback was fully provided in 17 of the 23 reviewed studies, was completely absent in 4, and was treated as an independent variable in only 2. From the 17 studies with feedback, only 2 failed to show testing effects, whereas from the 4 studies without feedback, only 1 study failed to show testing effects. Thus,

most studies found successful testing effects, independently of whether feedback was provided or not. The most appropriate manner to determine whether feedback enhance the effects of retrieval, however, is by comparing test conditions with and without feedback in the same experiment (i.e., treating feedback as an independent variable), as in Lipko-Speed et al. (2014) and Vojdanoska et al. (2010). From these 2 studies, 1 found testing effects only when feedback was provided (Lipko-Speed et al., 2014) and 1 found testing effects regardless of whether feedback was provided or not (Vojdanoska et al., 2010). Thus, even though feedback has been shown to enhance testing effects in laboratory settings (e.g., Butler and Roediger, 2008, although see Adesope et al., 2017), it remains unclear whether it is advantageous in classroom settings. Notably, when feedback is provided for all experimental conditions, it becomes difficult to isolate the effects of retrieval from potential effects of feedback (Karpicke, 2017). Thus, as in Lipko-Speed et al. (2014) and Vojdanoska et al. (2010), future research should examine the necessity of using feedback in classroom settings by directly comparing conditions in which feedback is provided to conditions in which feedback is not provided.

Does the Current Applied Literature Sustain the Recommendation of Retrieval Practice in Schools?

The current literature suggests that retrieval practice can be a useful learning tool in educational settings. This is in consonance with the idea that tests can be recommended as a strategy for learning in actual educational contexts (Roediger and Pyc, 2012). Caution should be taken, however, when replacing other learning activities by retrieval practice. It is well-demonstrated by the reviewed studies that retrieval practice is significantly more beneficial than "shallow" learning strategies, as reread for example. Further research is still needed to elucidate whether retrieval practice is more beneficial than more "active" learning strategies, as concept mapping, for example.

Further Issues Concerning Study Materials, Time of Retention, Motivation, Collaborative Testing, and Individual Differences

The current review showed that testing effects can be yielded in classroom settings for different age ranges, for different study materials, and when different intervals between study and retrieval practice are adopted. Regarding study materials, positive testing effects were found for topics as diverse as 7th grade science and history (McDermott et al., 2014), high school probability calculations (Dirkx et al., 2014), and college level anatomy and physiology (Dobson and Linderholm, 2015). Finally, the reviewed studies showed that the interval between study and retrieval practice is not determinant for the testing effects to occur in classroom settings. Studies using both short (e.g., Vojdanoska et al., 2010) and long study-retrieval intervals (e.g., Carpenter et al., 2009) successfully yielded long-term testing effects.

Another important issue for the application of retrieval practice in classroom contexts concerns the motivation of students to engage in such practice. Only 1 study manipulated this issue directly (Kibble, 2007), showing that students performed multiple-choice quizzes considerably more frequently when their engagement in quizzes was incentivized with small percentages of the final grades. The result of this study is very promising, and future research will be needed to examine whether the reward schedules adopted by those authors can be successfully applied to different types of tests (e.g., short-answer).

Although the remaining studies did not manipulate motivation, several included materials that were actually covered in final exams or unit exams, a fact that can perhaps enhance the motivation of students to engage in retrieval practice. As can be seen in the **Table 1** (column "Materials for actual exams?"), in all studies using multiple-choice tests, participants practice retrieval with contents that would be queried in later actual exams. Conversely, in only one study using short-answer tests participants practiced retrieval with contents queried in actual exams. It is not clear whether this is a contributing factor for the absence of testing effects in a subset of the short-answer studies. Future research should investigate this issue more closely, especially because short-answer tests are supposedly more effortful than more "passive" learning techniques (Pyc and Rawson, 2009), and motivation may be an important factor for the engagement of students in such type of test.

Only two studies examined whether performing tests collectively is more beneficial than performing tests individually. In both, undergraduate students were tested in psychology classes. In one of them, fill-in-the-gaps tests were equally beneficial when performed individually or in groups (Vojdanoska et al., 2010), while in the other, fill-in-the-gaps and multiple-choice tests were more beneficial when performed collectively than individually (Cranney et al., 2009). Thus, using retrieval practice collectively seems to be an effective alternative to improve learning in college level classes. It is at least as beneficial as retrieval practice performed individually.

Finally, it is important to note that only two of the reviewed studies approached the issue of individual differences. They showed that IQ differences were not determinant for the strength of testing effects in 3rd grade children (Jaeger et al., 2015), and that only students exhibiting high performance in prior course exams were benefited by short-answer tests in an introductory biology class (Carpenter et al., 2016). The lack of applied retrieval practice studies observing individual differences is surprising, given the importance of such differences in educational contexts. This deficiency in the literature, nonetheless, probably reflects a general lack of retrieval practice studies approaching individual differences. The only studies examining this issue in laboratory settings showed that retrieval can be beneficial for individuals with different neurological conditions (Sumowski et al., 2010a,b), for individuals from various age ranges (Tse et al., 2010), for individuals with low general-fluid intelligence and low episodic memory capacities (Brewer and Unsworth, 2012), for individuals with high working memory capacities and low test-anxiety scores (Tse and Pu, 2012), and showed that retrieval practice is not particularly beneficial for individuals with attention-deficit hyperactivity disorder (Dudukovic et al., 2015). Thus, examining the benefits of retrieval practice taking into account individual differences is certainly a very promising avenue for future research, both in applied and laboratory contexts.

Summary of Main Findings

As discussed in the Introduction section, our main goal was to approach a set of questions concerning the application of retrieval practice to educational contexts. Below, these questions are reiterated and followed by answers we have drawn from the findings of the reviewed studies.

(1) *Is the testing effect replicable in real educational settings?* Yes, since the majority of the reviewed articles found positive answers for this question.

(2) *Does the type of test matter?* The type of test matters. Multiple-choice and fill-in-the-gaps were shown to be highly beneficial for students. Short-answer tests, on the other hand, were not so consistently beneficial. These findings, however, should be considered in light of the type of control conditions used in each experiment and in light of the age of the participants (see questions 3 and 4 below).

(3) *Does the age of the students matter?* Age matters when the type of test is considered. That is, children in the 6th grade and below were not benefited by short-answer tests (Goossens et al., 2016), unless feedback was provided (Lipko-Speed et al., 2014). Children of this age range, however, were consistently benefited by multiple-choice (Roediger et al., 2011) and fill-in-the-gaps (Jaeger et al., 2015) tests.

(4) *Is retrieval practice advantageous in comparison to "stronger" control conditions?* In classroom settings, retrieval practice was shown to be more beneficial for learning than particularly "superficial" learning strategies, as reread; and shown to be more beneficial than performing no-activity whatsoever (i.e., the "no-activity" condition). However, the benefits of retrieval practice disappeared when tests were contrasted with more "active" learning strategies, as concept mapping (Burdo and O'Dwyer, 2015) or with the copying written materials (e.g., Goossens et al., 2016). Thus, given the reviewed research, retrieval practice is more beneficial than "weak" or "superficial" learning strategies, but it does not seem to be more beneficial than "stronger" learning strategies.

(5) *Does corrective feedback enhances the benefits of retrieval practice?* Considering the reviewed literature, it is not possible to infer that feedback is significantly beneficial for retrieval practice in educational contexts. Retrieval practice can definitely be beneficial without feedback, but too few (and heterogeneous) studies compared conditions with and without feedback to elucidate the real contribution of adding feedback to testing.

(6) *Is the current applied literature substantial enough to instill the recommendation of retrieval practice in school environments?* Yes, but caution should be taken concerning the activities that will be substituted by retrieval practice, since so far it is only possible to conclude that retrieval practice is more beneficial than reread or "no-activity" in actual educational contexts.

CONCLUSIONS

The reviewed articles show that testing effects can be in general successfully reproduced in classroom settings, with typical classroom materials. The types of control conditions and the age of the participants seem to have an important role in the success of retrieval practice, however. That is, retrieval practice was not reliably beneficial when compared to concept mapping or activities involving the copy of written materials, and when short-answer tests were used with children. These findings are important for educational purposes, especially if the adoption of retrieval practice implies the abandonment of other class activities. The question then is whether retrieval practice is more beneficial than the activities it may potentially replace. Future research should explored this issue by comparing retrieval practice to activities typically administered in class, instead of comparing retrieval practice with repeated reading or "no-activity." Thus, although considerable work should be done to elucidate these issues, the reviewed studies show that retrieval practice in the form of multiple-choice and fill-in-the-gaps tests are a promising learning strategy to be used in classroom settings.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

FUNDING

This work was supported by the Instituto Nacional de Ciência e Tecnologia sobre Comportamento, Cognição e Ensino, with support from the Brazilian National Research Council (CNPq, Grant No. 465686/2014-1), the São Paulo Research Foundation (Grant # 2014/50909-8), and the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES, Grant No. 88887.136407/2017-00). This work was also supported by the Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG, Grant No. APQ- 01174-14), and by the Brazilian National Research Council (CNPq, Grant No. 448537/2014-1).

REFERENCES

Adesope, O. O., Trevisan, D. A., and Sundararajan, N. (2017). Rethinking the use of tests: a meta-analysis of practice testing. *Rev. Educ. Res.* 87, 659–701. doi: 10.3102/0034654316689306

Balch, W. R. (1998). Practice versus review exams and final exam performance. *Teach. Psychol.* 25, 181–184. doi: 10.1207/s15328023top2503_3

Bangert-Drowns, R. L., Kulik, J. A., and Kulik, C. L. C. (1991). Effects of frequent classroom testing. *J. Educ. Res.* 85, 89–99. doi: 10.1080/00220671.1991.10702818

Batsell, W. R., Perry, J. L., Hanley, E., and Hostetter, A. B. (2016). Ecological validity of the testing effect. *Teach. Psychol.* 44, 18–23. doi: 10.1177/0098628316677492

Bjork, R. A. (1988). "Retrieval practice and the maintenance of knowledge," in *Practical Aspects of Memory II*, eds M. M. Gruneberg, P. E. Morris, and R. N. Sykes (London: Wiley), 396–401.

Brewer, G. A., and Unsworth, N. (2012). Individual differences in the effects of retrieval from long-term memory. *J. Mem. Lang.* 66, 407–415. doi: 10.1016/j.jml.2011.12.009

Burdo, J., and O'Dwyer, L. (2015). The effectiveness of concept mapping and retrieval practice as learning strategies in an undergraduate physiology course. *Adv. Physiol. Educ.* 39, 335–340. doi: 10.1152/advan.00041.2015

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

Butler, A. C., Karpicke, J. D., and Roediger, H. L. (2008). Correcting a metacognitive error: Feedback increases retention of low confidence correct responses. *J. Exp. Psychol. Learn. Memory Cogn.* 34, 918–928. doi: 10.1037/0278-7393.34.4.918

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

Callender, A. A., and McDaniel, M. A. (2009). The limited benefits of rereading educational texts. *Contemp. Educ. Psychol.* 34, 30–41. doi: 10.1016/j.cedpsych.2008.07.001

Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: the benefits of elaborative retrieval. *J. Exp. Psychol. Learn. Memory Cogn.* 35, 1563–1569. doi: 10.1037/a0017021

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

Carpenter, S. K., Lund, T. J., Coffman, C. R., Armstrong, P. I., Lamm, M. H., and Reason, R. D. (2016). A classroom study on the relationship between student achievement and retrieval-enhanced learning. *Educ. Psychol. Rev.* 28, 353–375. doi: 10.1007/s10648-015-9311-9

Carpenter, S. K., Pashler, H., and Cepeda, N. J. (2009). Using tests to enhance 8th grade students' retention of U.S. history facts. *Appl. Cogn. Psychol.* 23, 760–771. doi: 10.1002/acp.1507

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

Craik, F. I. M., and Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *J. Exp. Psychol. General* 104, 268–294. doi: 10.1037/0096-3445.104.3.268

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

Daniel, D. B., and Broida, S. (2004). Using web based quizzing to improve exam performance: lessons learned. *Teach. Psychol.* 31, 207–208. doi: 10.1207/s15328023top3103_6

Davidson, M. C., Amso, D., Anderson, L. C., and Diamond, A. (2006). Development of cognitive control and executive functions from 4 to 13 years: evidence from manipulations of memory, inhibition, and task switching. *Neuropsychologia* 44, 2037–2078. doi: 10.1016/j.neuropsychologia.2006.02.006

Dirkx, K. J. H., Kester, L., and Kirschner, P. A. (2014). the testing effect for learning principles and procedures from texts. *J. Educ. Res.* 107, 357–364. doi: 10.1080/00220671.2013.823370

Dobson, J. L., and Linderholm, T. (2015). Self-testing promotes superior retention of anatomy and physiology information. *Adv. Health Sci. Educ.* 20, 149–161. doi: 10.1007/s10459-014-9514-8

Dudukovic, N. M., Gottshall, J. L., Cavanaugh, P. A., and Moody, C. T. (2015). Diminished testing benefits in young adults with attention-deficit hyperactivity disorder. *Memory* 23, 1264–1276. doi: 10.1080/09658211.2014.977921

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., and Willingham, D. T. (2013). Improving students' learning with effective learning techniques: promising directions from cognitive and educational psychology. *Psychol. Sci. Public Interest* 14, 4–58. doi: 10.1177/1529100612453266

Eisenkraemer, R. E., Jaeger, A., and Stein, L. M. (2013). A systematic review of the testing effect in learning. *Paideia* 23, 397–406. doi: 10.1590/1982-43272356201314

Goossens, N. A. M. C., Camp, G., Verkoeijen, P. P. J. L., Tabbers, H. K., Bouwmeester, S., and Zwaan, R. A. (2016). Distributed practice and retrieval practice in primary school vocabulary learning: a multi-classroom study. *Appl. Cogn. Psychol.* 30, 700–712. doi: 10.1002/acp.3245

Jaeger, A., Eisenkraemer, R. E., and Stein, L. M. (2015). Test-enhanced learning in third-grade children. *Educ. Psychol.* 35, 513–521. doi: 10.1080/01443410.2014.963030

Kang, S. H. K., McDermott, K. B., and Roediger, H. L. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *Eur. J. Cogn. Psychol.* 19, 528–558. doi: 10.1080/09541440601056620

Karpicke, J. D. (2012). Retrieval-based learning: active retrieval promotes meaningful learning. *Curr. Dir. Psychol. Sci.* 21, 157–163. doi: 10.1177/0963721412443552

Karpicke, J. D. (2017). "Retrieval-based learning: a decade of progress,"in *Cognitive psychology of memory*. Learning and memory: a comprehensive reference, Vol. 2, eds J. H. Byrne, J. T. Wixted, (Oxford: Academic Press), 487–514. doi: 10.1016/B978-0-12-809324-5.21055-9

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

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

Kibble, J. (2007). Use of unsupervised online quizzes as formative assessment in a medical physiology course: effects of incentives on student participation and performance. *Adv. Physiol. Educ.* 31, 253–260. doi: 10.1152/advan.00027.2007

Kornell, N., Rabelo, V. C., and Klein, P. C. (2012). Tests enhance learning – Compared to what? *J. Appl. Res. Mem. Cogn.* 1, 257–259. doi: 10.1016/j.jarmac.2012.10.002

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

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

Lipko-Speed, A., Dunlosky, J., and Rawson, K. A. (2014). Does testing with feedback help grade-school children learn key concepts in science? *J. Appl. Res. Mem. Cogn.* 3, 171–176. doi: 10.1016/j.jarmac.2014.04.002

Lyle, K. B., and Crawford, N. A. (2011). Retrieving essential material at the end of lectures improves performance on statistics exams. *Teach. Psychol.* 38, 94–97. doi: 10.1177/0098628311401587

McDaniel, M. A., Agarwal, P. K., Huelser, B. J., McDermott, K. B., and Roediger, H. L. (2011). Test-enhanced learning in a middle school classroom: the effects of quiz frequency and placement. *J. Educ. Psychol.* 103, 399–414. doi: 10.1037/a0021782

McDaniel, M. A., Thomas, R. C., Agarwal, P. K., McDermott, K. B., and Roediger, H. L. (2013). Quizzing in middle-school science: successful transfer performance on classroom exams. *Appl. Cogn. Psychol.* 27, 360–372. doi: 10.1002/acp.2914

McDermott, K. B., Agarwal, P. K., D'antonio, L., Roediger, I. I. I., H. L., and McDaniel, M. A. (2014). Both multiple-choice and short-answer quizzes enhance later exam performance in middle and high school classes. *J. Exp. Psychol. Appl.* 20:3. doi: 10.1037/xap0000004

Mulligan, N. W., and Picklesimer, M. (2017). Attention and the testing effect. *J. Exp. Psychol. Learn. Memory Cogn.* 42, 938–950. doi: 10.1037/xlm0000227

Novak, J. D. (2010). Learning, creating, and using knowledge: concept maps as facilitative tools in schools and corporations. *J. e-Learn. Knowl. Soc.* 6, 21–30. doi: 10.4324/9780203862001

Pan, S. C., and Rickard, T. C. (2018). Transfer of test-enhanced learning: meta-analytic review and synthesis. *Psychol. Bull.* 144, 710–756. doi: 10.1037/bul0000151

Pashler, H., Rohrer, D., Cepeda, N. J., and Carpenter, S. K. (2007). Enhancing learning and retarding forgetting: choices and consequences. *Psychon. Bull. Rev.* 14, 187–193. doi: 10.3758/BF03194050

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

Pyc, M. A., and Rawson, K. A. (2009). Testing the retrieval effort hypothesis: does greater difficulty correctly recalling information lead to higher levels of memory? *J. Mem. Lang.* 60, 437–447. doi: 10.1016/j.jml.2009.01.004

Ramraje, S. N., and Sable, P. L. (2011). Comparison of the effect of post-instruction multiple-choice and short-answer tests on delayed retention learning. *Australas. Med. J.* 4, 332–339. doi: 10.4066/AMJ.2011.727

Roediger, H. L., Agarwal, P. K., McDaniel, M. A. and McDermott, K. B. (2011). Test-enhanced learning in the classroom: long-term improvements from quizzing. *J. Exp. Psychol. Appl.* 17, 382–395. doi: 10.1037/a0026252

Roediger, H. L., and Karpicke, J. D. (2006a). The power of testing memory: basic research and implications for educational practice. *Perspect. Psychol. Sci.* 1, 181–210. doi: 10.1111/j.1745-6916.2006.00012.x

Roediger, H. L., and Karpicke, J. D. (2006b). Test-enhanced learning: taking memory tests improves long-term retention. *Psychol. Sci.* 17, 249–255. doi: 10.1111/j.1467-9280.2006.01693.x

Roediger, H. L., and Pyc, M. A. (2012). Inexpensive techniques to improve education: applying cognitive psychology to enhance educational practice. *J. Appl. Res. Mem. Cogn.* 1, 242–248. doi: 10.1016/j.jarmac.2012.09.002

Rowland, C. A. (2014). The effect of testing versus restudy on retention: a meta-analytic review of the testing effects. *Psychol. Bull.* 140, 1432–1463. doi: 10.1037/a0037559

Schwieren, J., Barenberg, J., and Dutke, S. (2017). The testing effect in the psychology classroom: a meta-analytic perspective. *Psychol. Learn. Teach.* 16, 179–196. doi: 10.1177/1475725717695149

Smith, M. A., and Karpicke, J. D. (2014). Retrieval practice with short-answer, multiple-choice, and hybrid tests. *Memory* 22, 784–802. doi: 10.1080/09658211.2013.831454

Stenlund, T., Sundström, A., and Jonsson, B. (2016). Effects of repeated testing on short- and long-term memory performance across different test formats. *Educ. Psychol.* 36, 1710–1727. doi: 10.1080/01443410.2014.953037

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

Sumowski, J. F., Wood, H. G., Chiaravalloti, N., Wylie, G. R., Lengenfelder, J., and Deluca, J. (2010b). Retrieval practice: A simple strategy for improving memory after traumatic brain injury. *J. Int. Neuropsychol. Soc.* 16, 1147–1150. doi: 10.1017/S1355617710001128

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

Tse, C. S., and Pu, X. (2012). The effectiveness of test-enhanced learning depends on trait test anxiety and working-memory capacity. *J. Exp. Psychol. Appl.* 18:253. doi: 10.1037/a0029190

Vojdanoska, M., Cranney, J., and Newell, B. R. (2010). The testing effect: the role of feedback and collaboration in a tertiary classroom setting. *Appl. Cogn. Psychol.* 24, 1183–1195. doi: 10.1002/acp.1630

Wiklund-Hörnqvist, C., Jonsson, B., and Nyberg, L. (2014). Strengthening concept learning by repeated testing. *Scand. J. Psychol.* 55, 10–16. doi: 10.1111/sjop.12093

Yonelinas, A. P. (2002). The nature of recollection and familiarity: a review of 30 years of research. *J. Mem. Lang.* 46, 441–517. doi: 10.1006/jmla.2002.2864

Yonelinas, A. P., and Parks, C. M. (2007). Receiver operating characteristics (ROCs) in recognition memory: a review. *Psychol. Bull.* 133, 800–832. doi: 10.1037/0033-2909.133.5.800

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.