Can tests improve learning in real university classrooms?

Mathilde Lamotte, Marie Izaute, Céline Darnon

Can Tests Improve Learning in Real University Classrooms?

Mathilde Lamotte

Marie Izaute

Céline Darnon

Université Clermont Auvergne, LAPSCO - CNRS UMR6024

Abstract

Long-term memory of a stimulus is likely to increase when individuals are tested on this stimulus, as compared to when they only re-study it. This "testing effect" suggest that tests could be used in real classroom contexts to increase students' academic performance. However, real classroom contexts can be considered as evaluative contexts that might change the meaning and effects of tests. The present paper reviews existing evidence regarding the positive effect of testing on learning in real-class academic settings and on learning material that is part of the curriculum. The present review underscores that positive effects of testing can occur in real classroom contexts, but points to the fact that such positive effects are reduced when tests are used as grading and selection tools rather than as learning tools. The review also points to important challenges that will have to be addressed in future research examining the testing effect in real classroom contexts.

Keywords: Testing, learning, classroom, University, retrieval practice.

General Audience summary

Many research has documented the existence of a "testing effect": when students are tested on the to-be-remembered information, they are likely to memorize it better than when they only restudy it. This suggests that tests could be used in real classroom contexts to increase students' academic performance. However, most of this research has been conducted in a lab context. Real classroom contexts can be considered as evaluative contexts that might change the meaning and effects of tests. The present paper reviews existing evidence regarding the positive effect of testing on learning in real-class academic settings and on learning material that is part of the curriculum. The review underscores that positive effects of testing can occur in real classroom contexts but also highlights important features regarding the implementation of tests in classrooms that can prove particularly useful for teachers. For example, depending on the time teachers can allocate to testing students during class time, they can use either multiple choice or short answer questions. If teachers prefer not to allocate class time to testing, an option could be to implement their tests outside of the classroom, such as with e-learning devices. Finally, the present review point to the fact that teachers should include corrective feedbacks after the test questions, particularly because corrective feedback should increase the formative value of tests and, consequently, increase the likelihood that students will perceive these tests as learning tools rather than selection tools, with further benefits for learning.

Can Tests Improve Learning in Real University Classrooms?

The number of students who enter higher education but drop out before graduating is particularly high and alarming in most Western countries. In France, for example, only 27% of first-year students obtain their bachelor's degree three years later (M.E.S.R., 2013). In the United States, approximately 30% of college freshmen drop out before their sophomore year (Ginder, Kelly-Reid, & Mann, 2017), and similar rates are observed in most Western countries (OECD, 2013). This drop-out rate is particularly alarming and strongly calls into question the teaching methods practiced in most higher education courses.

The question of how to increase learning and long-term retention has interested psychologists for years. Several practices have been examined by cognitive and social psychologists (for reviews, see Dunlosky et al., 2013; Yeager & Walton, 2011). Among these practices, retrieval practice, via the use of testing, is one of the practices that has produced a greater amount of research these last 30 years. Indeed, the beneficial effects of tests on learning is one of the most prolific topics in cognitive psychology, for instance, using "testing effect" and synonymous as keywords more than 88,000 results came through on PsycINFO. The conclusions of this abundant research are particularly consensual and straightforward: Compared to restudying, testing benefits learning (for reviews, see Roediger, Putnam, & Smith, 2011, and McDaniel, Roediger, & McDermott, 2007). Given their beneficial effects on learning, tests should be highly used in educational settings, including in higher education classes. However, quite surprisingly, thus far, most research in this area has been conducted in laboratory settings (for reviews see, Chan et al., 2018; Roediger & Karpicke, 2006b; Roediger et al., 2011; Rowland, 2014). In general, research conducted in laboratory settings documents an overall positive effect of testing (Adesope, Trevisan, & Sundararajan, 2017; Chan, Meissner, & Davis, 2018). These studies are conducted using a large variety of materials: pictures (e.g., Wheeler & Roediger, 1992), word pairs (e.g., Pashler, Rohrer, Cepeda, &

Carpenter, 2007; Pyc & Rawson, 2007, 2012), or parts of texts (e.g., Roediger & Karpicke, 2006a). However, these materials rarely relate to what students really have to learn to pass their exams at the end of the academic year. Some laboratory studies investigated the testing effect with material that can be considered as educationally relevant material (e.g. Einstein, Mullet, & Harrison, 2012; Nungester & Duchastel, 1982; Roediger & Karpicke, 2006a). The findings also bring support to the benefit of testing on learning.

Indeed, as compared with laboratory setting studies, relatively few studies have examined the effect of testing in real classroom contexts and with real learning content, especially in the university context. The use of research results obtained in the laboratory is essential for designing efficient pedagogical practices (Connolly, Keenan, & Urbanska, 2018; Dehaene, 2019). Yet the transposition of lab-based results into real classroom contexts is not straightforward (see Sotola & Crede, 2020 for a similar argument). In particular, real classroom contexts are likely to elicit various forms of motivations and goals that can significantly impact the way students apprehend the learning situation (Darnon, Poortvliet, & Dompnier, 2012; Huguet & Kuyper, 2008). The very meaning of "tests" is likely to differ in these two contexts, especially as succeeding on a test is much more important for students when this test is part of their academic curriculum than when it has no consequences on future pass/fail decisions (Crooks, 1988). Interestingly, in their review, Roediger and Karpicke (2006a) highlighted the fact that, in general, both teachers and students tend to consider tests as assessment devices rather than learning tools, with potential consequences on their effects on learning. This issue is important because research has identified that the significance attributed to a test (whether the test is used as a tool for learning or for comparison and selection) can significantly impact its effects (see Darnon et al., 2011). For example, in Smeding et al. (2013), students enrolled in a statistical class had to take a test presented as part of the learning process (mastery-oriented assessment) or as a way to compare students to

each other and select the best of them (selection-oriented assessment). The research showed that the selection-oriented tests tended to increase the socioeconomic status (SES) achievement gap whereas the mastery-oriented tests decreased this gap (see also Jury, Smeding, & Darnon, 2015; Souchal et al., 2014). These results suggest that, when tests are presented (or used) as selection and comparison tools, they threaten low SES students' identity (Croizet & Claire, 1998), a process that may annihilate the positive effects they could have on learning outcomes. In lab contexts, students are quite free from evaluative pressure, which is not the case in real classroom contexts, where tests are the means by which teachers assess, compare, and rank students' performance and make pass/fail decisions, particularly when the test is based on part of the curriculum rather than on less significant material, like images of pairs of words.

The purpose of the present manuscript is to review existing evidence regarding the testing effect in a real university classroom context, with the proper content of the current curriculum as the learning (and test) content and real exam grades as measures of performance.

Why Should Testing Increase Learning?

Several literature reviews and meta-analyses have been published throughout the years (e.g. Adesope, Trevisan, & Sundararajan, 2017; Chan et al., 2018; Moreira et al., 2019; Pan & Rickard, 2018; Rowland, 2014; Schwieren, Barenberg, & Dutke, 2017; Sotola & Crede, 2020) and have confirmed the general beneficial effects of testing on further memory test performance, as compared to other learning strategies, with an effect size comprised between Hedge's g = .50 (Rowland, 2014) and g = .61 (Adesope et al., 2017). For example, in a meta-analysis focusing on 61 studies investigating the effect of testing on subsequent learning, Rowland (2014) found an overall robust positive effect of testing on learning outcomes as compared with a restudy condition. Interestingly, contrary to students' beliefs (e.g., R. A.

Bjork, Dunlosky, & Kornell, 2013; Kornell & Son, 2009), in lab settings, tests have even proved more efficient for learning outcomes than the restudy of the same material. As an example, Zaromb and Roediger (2010) asked participants to study three lists of 50 words belonging to 30 categories during four sessions and compared three conditions. In one condition, participants studied the material during the 4 sessions. In another, they studied the material during the first 2 sessions and were then tested during the other two. Finally, in a third condition, they were tested during the 4 sessions. The results indicated that both test conditions significantly improved the proportion of correctly recalled words. As previously mentioned, such a positive effect of testing, as compared to restudy, has been documented several times with various materials, including pairs of words, prose passages, and maps (for a review, see Rowland, 2014).

This positive effect of testing stands on the fact that tests increase retrieval efforts. Indeed, when answering the test, students have to produce an effort to retrieve previously learned material from their memories. In addition to other factors which are also likely to contribute to the positive effects of testing on learning including motivation (e.g., Kang & Pashler, 2014) or additional exposure (Roediger & Karpicke, 2006b), this retrieval practice presents several benefits. First, it improves later retention on the same or related topics (e.g., Carpenter, 2012; Karpicke & Roediger, 2005; Wheeler & Roediger, 1992). Second, it allows learners to identify potential gaps in knowledge, which is likely to increase future acquisitions (e.g., Carpenter & Delosh, 2005, 2006; Hays, Kornell, & Bjork, 2013; Kornell, Hays, & Bjork, 2009; Roediger & Karpicke, 2006a), resulting in both direct and indirect benefits for further performance (see for instance Arnold & Mcdermott, 2013 for the test-potentiated learning effect; see also Wissman & Rawson, 2018 for recent replications and developments). In addition, during retrieval practice, elaborative memory traces are formed due to the activation of information related or not to the targeted information (Carpenter & Delosh,

2006). In other words, a network of connections between information would be created during retrieval practice that leads to reinforcing previously learned elements and facilitating the learning of new elements during subsequent learning sessions. Moreover, testing also facilitates the transfer of learning meaning that being tested on one type of learnt material or in a given context can further improve subsequent learning in other contexts (for a meta-analysis and review of this topic, see Pan & Rickard, 2018). In addition, retrieval practices are susceptible to generating "desirable difficulty" more than the mere study of the material. For instance, using either short ("easy condition") or longer ("difficult" condition) interstimulus intervals, Pyc and Rawson (2009) documented that the beneficial effect of testing on performance mostly occurred for the difficult stimuli.

In addition, one findings of this literature is that the higher the similarity between the original content and the final test content, the higher the testing effect (Chan et al., 2018). However, this result is not observed consistently. Indeed, Rowland's meta-analysis (2014) did not find such an effect of similarity on the magnitude of the testing effect. The test format (e.g., item-cued recall, free recall) does not seem to moderate the effect of testing, but a delay could, although in this review (Chan et al., 2018), only 10 samples included a long (longer than 24h) delay while all of the remaining 149 samples had a relatively short interval retention (< 24h). Adesope and colleagues (2017) also found that, irrespective of the features of the test (e.g., free recall, cued recall), all types of tests led to better learning outcomes than non-retrieval learning practices (e.g., restudying, filler, other activities). However, unlike Chan and colleagues, they identified larger effect sizes for the mixed test format and multiple choice question (MCQ) tests than with free-recall, cued-recall and short answer question (SAQ) tests. In particular, the strongest effect sizes emerged when test practices and the final exam were in the same format. Of great interest for teachers' practice, feedback provided after tests does not seem to impact the size of the testing effect. In both cases (with or without

feedback), the effect sizes remained relatively high (with feedback $g = 0.63$ or without $g = 0.60$). This discrepancy between Chan and colleagues' findings and those from Adesope and colleagues can also be explained by the fact that the former focused on the effect of testing on subsequent learning although the latter explored the effect of testing on previously learnt material. This is of great interest for our purpose. Indeed, in the context of academic learning, students' assessment usually follows periods of learning and revision.

However, as previously mentioned, most of the studies reviewed in these meta-analyses focused on laboratory experiments and were conducted with quite neutral materials. What about the test effect in real classroom contexts, in which tests are usually used as grading instruments and are thus susceptible to eliciting various forms of motivation, threats, and other related concerns not elicited in a lab setting?

Does Testing Improve Learning in Real Classroom Contexts?

The vast majority of research examining the test effect has been conducted in laboratory settings. However, some research has also examined this effect in real-classroom contexts and some reviews and meta-analysis have recently emerged on this issue. Schwieren et al.'s (2017) is one of them. This meta-analysis was conducted among psychology classes and confirmed the beneficial effect of testing on subsequent learning outcomes as compared to control conditions (i.e., no test or restudy), with a moderate effect size ($d = .56$). Interestingly, the only significant moderator of the testing effect that appeared in this meta-analysis was feedback; providing the correct answer after the test seemed to increase the beneficial effect of the test on the subsequent exam performances. Another recent and particularly relevant review highlighted an overall positive effect of retrieval practice to improve learning in real classroom settings (Moreira et al., 2019). In this study, incongruently with lab-settings research and predictions from literature, the advantage of SAQ over the MCQ is not clear, sustaining the idea that the type of tests can take a specific meaning in real

classroom contexts. Although their research was not specifically focused on University students, they conclude on the fact that the inherent characteristics of classroom settings might explain such inconsistencies.

Contrary to this previous review which was focused on all age groups, the purpose of the present paper is to fully review the research conducted in classroom contexts at a University level. Specifically, we focus on research meeting the following criteria: (1) articles were published in peer-reviewed journals; (2) participants were university students; (3) a test condition was compared to a control condition that did not involve a retrieval practice (see below); and (4) the material to be learned and the test content were related to the class curriculum. In real classroom contexts, it is unclear what should be considered an appropriate control group. Depending on the studies, in the control group, students either had to rewrite the test questions in an affirmative form, without answering them, read some notes from which the test items were created, or restudy the material. The common characteristic of these control groups is that they do not imply retrieval. Consequently, one of the criteria for the inclusion in this review was the presence of a control group in which retrieval practice was not encouraged but another type of revision was implemented (for a review of this question, see Kornell, Rabelo, & Jacobs Klein, 2012).

Searching the PsycINFO database with “test effect” or “testing effect” or “retrieval practice” or “retrieval effect” or “test-induced learning” or “backward-testing effect” as keywords, and refining to “university students” or “college students”, “learning” as keyword and “class” or “classroom”, we obtained 33 results. We also explored the references from the meta-analysis and the selected articles as well as references suggested by the referees of a previous version of this paper[1] leading to 20 additional articles. A review of titles, abstracts and methods led us to eliminate 23 irrelevant articles based on our inclusion criteria: eight

[1] We are very grateful to the referees for their suggestions.

implying elementary/middle/high school students rather than university students, five were lab and not in-class studies; five were not published in peer-review journals; three were either meta-analysis or reviews of articles elsewhere published; two did not compare test with no-test condition and for one of the mentioned, the material was also not part of the curriculum. Finally, 30 articles, reporting a total of 39 studies, met the criteria and were retained. The characteristics of these papers are summarized in Table 1. Overall, they implied a total of more than 6,000 participants.

The studies reported in the selected articles differed in terms of type of tests used, their number, the type of final exam (e.g., SAQ or MCQ), participants' level, content of the material to be learned, and the presence or not of post-test feedback. Despite these differences, in all of the studies, an overall beneficial effect of the use of test appeared in the subsequent performance. Indeed, in general, students who were in the test conditions outperformed those in the control (no-retrieval) conditions on subsequent performance tests. These studies are discussed with greater details below.

Testing Effect in Real Classroom Contexts: Review of Existing Evidence

In fourteen of the selected studies, the tests were directly implemented within classes and compared to in-class practices that did not involve retrieval. For example, Bjork and colleagues (E. L. Bjork, Little, & Storm, 2014) studied students attending a 10-week basic research methods course that was assessed with several graded tests throughout the semester and a final exam at the end. Students received MCQ tests about 3-4 days after the lecture. Then, 3-4 days after the MCQ tests, they received their grade without any discussion or correction of it. Finally, they all took the final exam, which consisted of 50 items, among which 15 were relevant for the analysis (5 previously tested, 5 related, and 5 baseline control—i.e., never tested before). The results indicated that both previously tested items and conceptually related but untested items were more successfully recalled than the baseline

items. Other studies, also using between-participants designs, showed a beneficial effect of being tested in-class as compared with no retrieval practice condition on the delayed memory outcome. For example, Balch (1998) highlighted that students who received in-class quiz outperformed those who just read again the learning material on the final exam. Similarily, Batsell et al., (2017) showed that students who completed MCQ-quiz in-class got better grade at final exams than students who only red the material. Moreover, their results demonstrated that whereas there is no difference between pre-exposed and new items for students who only red the material, those who were quizzed had better performance for both types of pre-exposed (identical and related but rephrased) items as compared with the new ones.

Leeming (2002) also directly implemented an "exam-a-day procedure" in a class and compared students' performance in this class with students' performance from a previous similar class (with no such procedure). Participants were 192 undergraduate students from a learning and memory class and an introductory psychology class. In the experimental class, they completed 22 to 24 exams during class time. All of these exams were graded. Each consisted of two short essay questions and five SAQs and was immediately followed by corrective feedback (i.e., correct answers). The results indicated that students in the experimental class (exam-a-day procedure) outperformed those in the control class (without exam) in their final grades. In addition, students in the experimental class were compared to students in a similar class taught by two other teachers who did not use the exam-a-day procedure but only administered three cumulative exams. The results indicated that students in the experimental class again outperformed those in the control condition on the final exam. These results provide evidence that the testing procedure is efficient for improving learning in a real classroom environment when students practice testing within class in—almost—every class. This result is also supported by Inouye's study (Inouye, Bae, & Hayes, 2017) in which two to four pop fill-in-blank quizzes were implemented within the lecture. Students had to

respond immediately on a whiteboard support. The study compared participants who received the same lecture and exam with or without any quiz. In addition, for the participants who were in the whiteboard condition, they compared items that were tested in-class to those which were not tested. In both cases, the quiz led to better performance on the final exams.

Using a within-subject design in an educational psychology class, Chang (2018) conducted an experiment with 33 participants to compare three learning conditions, corresponding to three sub-units of the class: study–study versus test–test versus control (lecture). One third of the learning phase consisted of 10 MCQs (test condition) and 10 statements (study condition). In both experimental conditions, participants received MCQ/statements before and after the lecture. The 10 MCQs from the test–test condition were rewritten as 10 statements for the study–study condition. Finally, 10 additional items were presented during the post-test that could be considered as the measure of performance as they accounted for the grade. The order of questions, statements, and answer options were counterbalanced across participants. The results showed a main effect of the learning condition on the exam performance. A post-hoc analysis revealed a benefit of the test–test condition compared to the study–study condition. The control condition was in between, but differed from both of the two conditions. These results support the positive effect of testing on learning outcomes: Comparing their own performances, participants performed better when they were previously tested than when they restudied the learning material.

Also using a within subject design, Shapiro and Gordon (2012) compared three conditions: participant either received in-class quiz, do nothing more than the usual lecture, or saw the slides emphasized when relevant for the subsequent exam. The results showed that the performance for the previously tested items was better than for the not tested items and for the items related to emphasized slides. This finding is interesting as it tends to show that alternatives to restudy or re-reading existed but did not seem to be efficient to improve

students' learning. In the same line, the study from Dobson and Linderholm (2014) aimed at comparing three learning conditions. Across the semester, 125 students had to work on three different topics, each one randomly assigned to one of the learning conditions (re-reading vs. taking note while re-reading vs. testing). The results confirmed the beneficial effect of testing over both other conditions as the exam related to the tested topic led to better performance than the others.

In another study conducted with 140 undergraduate psychology students, Khanna (2015) compared three between-participant learning conditions: graded tests versus non-graded tests versus no quizzes (control condition). In the two former conditions, participants took six quizzes (1 final exam and 5 MCQ tests) that were either graded or non-graded. In the no-quiz condition, they only received the quiz that corresponded to the final exam. The results indicated that students in both test conditions showed increased performance on the final exam compared to the control condition. Moreover, performance was better in the non-graded than the graded condition. These results bring important insights regarding the significance of tests in such contexts. Indeed, as suspected, in real classroom contexts, where tests can be perceived as either learning tools or selection tools (Darnon et al., 2011; Souchal et al., 2014), non-graded tests seem to be more efficient than graded tests for increasing learning. However, it is worth noting, despite obvious, that the students had to eventually take the tests to benefit from it. Indeed, Trumbo and colleagues (Trumbo, Leiting, McDaniel, & Hodge, 2016) showed that whether students were assigned to required versus optional quiz conditions, the participants in the required quiz condition, meaning those who actually did practice testing, obtained better performance on the exams that those in the optional condition.

In a slightly different perspective, but relevant to our purpose, Carpenter, Rahman, and Perkins (2018) conducted a study with 230 psychology students. They introduced both pre-questions (i.e., questions that related to upcoming learning material) and post-questions (i.e.,

questions related to learned material). The dependent variable was the performance on an exam one week later. This study compared between questions that were previously tested twice (pre- and post-question condition) versus questions tested once (post-question condition) versus questions never tested. The results showed that students performed better on the questions that were tested (both twice and once) than with the "never tested" questions.

Taken together, the results of the studies discussed thus far support the idea that tests are efficient tools to be implemented in class to improve students' learning. However, the remaining question is whether this main positive effect of tests is qualified by certain moderators. In the following section, we describe studies that explore this issue.

Does the Type of Test Moderate the Effect of Testing in Real Classroom Settings?

In lab experiments, the debate regarding the effect of the type of test (i.e., MCQ or SAQ) is still vivid (J. D. Karpicke, 2017; Little, Bjork, Bjork, & Angello, 2012). This issue is of particular interest within classrooms because MCQs are probably easier to build and correct than SAQs. Consequently, MCQs could be more easily implemented by teachers than SAQs. In a series of four experiments conducted with 588 university students enrolled in a research method course, Foss and Pirozzolo (2017) compared the performance on a final exam for previously tested concepts versus untested concepts. In this research, the effect of the nature of the tests (i.e., MCQs versus SAQs) and that of the frequency of the test were also considered. All participants took the same final exam, which was composed of half MCQs and half SAQs. Among these items, half were new, and the other half corresponded to previously tested items. It is worth noting that part of the items of the exam were a flipped version of the previously tested MCQ/SAQ items. Overall, the results supported that the previously tested items resulted in a better performance than the new items—an effect that occurred for both types of questions and for both the flipped items and the previously tested items. These results provide an important replication of the testing effect in a real classroom

context but also suggest that the type of question did not seem to be crucial, as this positive effect of testing was observed both for MCQs and SAQs. Similarly, comparing two types of homework assignments, involving SAQ and MCQ, on the exam performance, Butler et al. (2014), found that practicing test led to better grades than not being tested whatever the question format. Lyle and Crawford (2011) compared 144 students who, depending on the condition, had to practice testing or not. The practice exercises were two to six items SAQ and fill-in-blank but the exam were always MCQ. In this study, test led to better exam performance despite the fact that the test and the exam format of the questions were different.

However, other research lead to different conclusions. For example, Butler and Roediger compared in a simulated classroom the format of questions (SAQ vs. MCQ) and the task (restudy vs. test vs. no activity). Their results showed a benefit of SAQ over the MCQ and of the testing over the restudy task over no-activity. Moreover, the interaction revealed that being tested with SAQ improved further performance as compared with restudy but there is no difference between the latter and being tested with MCQ despite both test format question conducted to better performance than no activity. Another recent study (Greving & Richter, 2018) involving 137 psychology students found similar results with a difference between MCQ and SAQ tests, in favour of SAQ. In this study, all students were randomly assigned to one of the three “practice session” conditions (SAQs vs. MCQs vs. restudy). The practice sessions took place during the last 10 minutes of 7 lectures (out of 10). During these, the participants received 8 SAQs, 8 MCQs, or 8 statements all rewritten from the lesson material. The first test took place a week after the tenth and last lecture, the second at the beginning of the first lecture of the following semester, and the third one at the end of the very last lecture of the following semester. The results suggested that SAQs but not MCQs increased performance when compared to the restudy condition. In other words, in this context involving testing during class that was not compulsory, only SAQs seemed to

improve learning outcomes. The discrepancy between Foss and Pirozzolo's (2017) results and Greving and Richter's (2018) might be due to the fact that the former did not compare the three conditions directly. Nonetheless, more investigation is needed before clear conclusions can be drawn on the possibility of a greater testing effect for SAQs when compared to MCQs in classroom context.

Do the Tests Have to be Taken During Class Time to Improve Learning?

Although tests are efficient tools for learning, they take up some class time. In addition, the time students have to study on their own is usually more important than the time they have in front of a teacher in the classroom. Thus, the use of tests outside the classroom could represent a promising perspective to implement in students' habits. Several possibilities occur for the teachers. Among these options, authors have examined the use of test from textbooks, the use of online homework assignments, either compulsory or not, the use of laboratory assignments.

In Welch's (2019) study, the tests came from the use of a textbook. The participants were 544 students enrolled in an introductory psychology course based on a textbook. Those in the control condition received a lecture but did not use the textbook. In the learning curve condition, students received a lecture and had to answer questions. They also received feedback on correctness. Indeed, they could either directly access the correct response or the relevant page of the textbook. Finally, in the learning practice condition, students received the lecture plus the same question as in the learning curve condition, but they also received two additional MCQs. All participants followed the course and were assigned randomly to conditions (no textbook, LC, LP). In addition, the responses to the MCQ in both of the experimental conditions (i.e., learning curve and learning practice conditions) accounted for 10% of the final grade. The results showed that learning curves alone did not lead to better performance than the control condition, although the learning practice condition did. These

results support the idea that testing remains efficient for improving learning even when the tests are not taken during class time.

In a similar way, using a within-subjects design, Kelley and colleagues (2019) tested whether the use of an online tool supporting retrieval practices ("PeerWise") could increase students' final grades. In this study, 40 psychology students completed online assignments from the PeerWise tool that consisted of generating at least eight questions and answering questions from other students for each chapter throughout the semester before due dates. Completion of the PeerWise assignments accounted for at most 4% of students' final grade. The final exam consisted of 60 MCQs and 12 SAQs. Some of these items were related to the tested/generated questions on PeerWise while others were new. All students received feedback about the correctness of their response and the page number to find the relevant information within the textbook. The results highlighted that previously tested items were more accurately answered than non-tested items. Interestingly, the results also showed that the more students used the PeerWise tool, the better their performance was on the final exam, particularly with MCQs (but not significantly with SAQs). Four other studies explored whether or not and how online testing can improve students' learning. Butler et al. (2014) submitted 40 students to two conditions of online homework assignments: half of the participants did the assignments as usual whereas the other half received additionally repeated retrieval practices that consisted in ten to twenty items, first SAQ then MCQ. Congruently with Kelley's results, students who received the testing practice got better performance on the exam than the students who only received the usual online assignment.

Interestingly, Grimstad and Grabe (2004) invited students to voluntarily practice testing as part of an online homework assignment. They compared "users", namely, students who did use the pool of questions to be tested (answered at least 50 out of 100 questions), to "non-users" (i.e., students who did not take the tests). They showed that "users" who

practiced testing obtained better performance on the exams than “non-users”. This finding is in line with Trumbo et al. (2016) but is also promising as it brings to light that students could, on their own, when the possibility is encouraged, being tested to improve their achievement. Another study with a quite similar design also brings support to the interest of online testing on students’ learning. Indeed, Johnson and Kiviniemi (2009) showed that tested items led to greater performance on the subsequent exam than non-tested items. In addition, a positive correlation between the number of completed tests through semester and the average score on the exams and the course grade was obtained: the more students completed tests, the better their scores. On the contrary, Bell and colleagues (Bell, Simone, & Whitfield, 2015) did not find such effect of online testing on subsequent in-class quiz performance neither on the exam scores. Thus, the context in which the online testing occur or the fact that it is followed by additional in-class testing might affect its effect on students’ learning.

Wiklund-Hörnqvist et al., (2014) also used computer lab sessions, in addition to the assigned readings and lectures in their classes. This study was conducted with 83 undergraduate students enrolled in a cognitive psychology course that included a voluntary learning lab session as well as the usual class. During these lab sessions, depending on the condition to which they were randomly assigned, participants either received repeated tests with feedback or restudied the learning content of the class. The main result of this study indicated that, after 18 days of delay, participants in the testing condition performed better than those in the restudy condition.

Similarly, McDaniel and colleagues (2007) compared tests to an exposure-only condition using an online procedure. Thirty-four students enrolled in a brain and behavior course took part in the experiment. The weekly tests included both MCQs and SAQs and were followed by feedback. In the control condition, students were only exposed to the target material. The results indicated that the weekly tests, but not the additional reading, improved

students' performance on the exams. Interestingly, in this experiment, these benefits seemed to be stronger for SAQs than for MCQs.

Taken together, the findings of these studies using tests within the curriculum but outside the classroom highlighted that the testing effect also appears when students are prompted to use tests on their own. In other words, using tests both in class and outside the classroom could improve students' learning and achievement. This issue has important practical implications. Indeed, teachers may consider that they do not have enough time during their class to have students work on some tests. These results suggest that, if they lack time, a relevant option for them could be to have these compulsory tests answered outside of the classroom, such as on online platforms. In addition to saving teaching time, such an approach would also space the learning and practice, as spaced practice is usually more efficient for learning than massed practice (e.g., Cepeda et al., 2006).

Exploring Other Ways to Increase the Benefits of Testing in Real Classroom Contexts

In a recent article, Eastridge and Benson (2020) compared two conditions of testing: individual testing versus collaborative testing (N = 129). Although the lack of a control group as we defined it at the beginning as a no-test condition in this experiment prevents us from drawing clear conclusions on the beneficial effects of testing on learning, the results of this experiment point to an interesting feature of testing that should probably be more deeply investigated in future research. In Eastridge and Benson's experiment, the tests were either taken collectively or individually. More precisely, students received four tests throughout the semester. These tests were either all taken individually (authors' control group, previous year students) or both individually (two tests) and collaboratively (two others). The results showed that students who completed the tests collaboratively obtained higher scores than those who completed the tests individually.

Previous works from Cranney and colleagues also support this finding. In two studies (Cranney et al., 2009), they showed that when the test activity was completed collaboratively, students performed better on the final exam than when the testing activity was completed individually. Moreover, although the individual testing led to better performance as compared with no activity, there was no difference between being tested individually or restudy. In other words, in this study, the beneficial effect of testing only occurred when students completed the test collaboratively rather than individually. In another study (Vojdanova, Cranney, & Newell, 2010), participants were either randomly assigned to a collaborative testing or an individual testing conditions. The results highlighted that on the initial exam, again, collaborative testing led to better performance than individual testing.

Recently, Thomas and colleagues (2020) implemented an in-class study with 45 students, using a mixed design. Students were assigned to a condition of restudy vs. quiz vs. quiz + feedback, and a condition of intermediate exam (collaborative vs. individual). Conditions were counterbalanced across the semester. The results showed that both type of testing led to better performance than restudy. Moreover, a greater improvement in performance occurred on related items rather than the identical ones for both types of exam. In accordance with the results from Cranney and colleagues and Eastridge and Benson, students who were tested collaboratively during the intermediate exam were more successful at the final exam than those who were tested individually on identical items.

These results make sense considering that collaborative practices usually reduce performance goals and social comparison (Buchs, Butera, & Mugny, 2004; Butera & Buchs, 2019; Nichols, 1996). We believe that, in such collaborative contexts, tests are less likely to be perceived as selection tools and more likely to be perceived as learning devices than in competitive or individualistic contexts.

Another promising result, yet to need further explorations, comes from Dobson and Linderholm (2014). In this study, after being affected to one condition of learning strategy, all students received full information regarding the beneficial effect of being tested on performance. Then, the performance the students who received such information obtained at the exams was compared to that obtained by students who attended the same course but did not received such information. Students who received this information had better grades on the final exams that students who did not. This effect is encouraging and should be further explored to understand more in depth if the impact of in-class testing can be enhanced by introducing information about how useful testing is for students' learning.

Testing in Real Classroom Contexts: Perspectives for Future Research

Summary of the Main Findings and Implications for Practice

The studies reviewed in the present paper were all conducted in real classroom contexts and used curriculum-based learning materials, tests, and exams. The existing evidence is based on a vast variety of methods, but overall, the findings support those previously obtained in laboratory settings: Compared to control groups (with no retrieval practice), in real classroom settings, testing increases subsequent memory performance. In the same vein, the format of the performance test is highly variable from one study to another—an inconsistency also identified in research conducted in lab settings (Chan et al., 2018; Rowland, 2014). Despite this inconsistency, the findings are quite stable and support that, irrespective of their format or length, all types of tests seem to benefit learning as compared with restudy condition. Of course, one cannot exclude this overall positive conclusion regarding the effect of testing on learning to be at least partially due to a publication bias and the fact that usual practices in the field (e.g., determining the relevance of results by relying only on significance testing) encourage false-positive. Indeed, the current process of publication tends to foster studies that produce significant results (Ferguson & Brannick,

2012; Francis, 2012; Świątkowski & Dompnier, 2017). However, it is important to note that Schwieren et al. (2017) did not find any indication of a publication bias in the data set they explored.

These findings have important implications for practice. Indeed, as noted in the introduction of the present paper, the large number of students who fail their exams challenges the pedagogical practices used in most university classes. Of course, this large rate of failure can be explained by several factors not related to teachers' practices (e.g., Jeno, Danielsen, & Raaheim, 2018; Stephens et al., 2015). However, the results reviewed and discussed in the present paper suggest that, all things being equal, the use of tests throughout the semester can significantly increase learning. As such, we believe such an approach deserves to be used in higher education class contexts in order to reduce the likelihood of students dropping out (Kift, 2015; OECD, 2013).

Identifying conditions for tests to increase learning

Although the results of the present review support that, in general, testing benefits learning, they also point to the fact that more research is needed regarding the existence of potential moderators of this positive effect in real classroom settings. Notably, based on the reviewed studies, it is unclear whether some types of questions produce higher learning gains than others. In the reviewed articles, ten used exclusively MCQs, two used exclusively SAQs, one did not give enough information regarding the type of tests and the others used a mixed of SAQ and MCQ or fill-in-blank questions. Except for Greving and Richter (2018) and Butler and Roediger (2007), who found an advantage of SAQs compared to restudy but not of MCQs compared to restudy, all the others found the testing effect with both types of questions. Foss and Pirozzolo (2017) sought specifically to compare the effects of two types of test questions on learning. They observed, as well as Thomas et al., (2020), that both types of tests increased learning. In other words, in line with the findings obtained in lab settings (Adesope et al.,

2017; Chan et al., 2018), these results suggest that testing leads to better performance on learning outcomes, whatever the test type (MCQ or SAQ). However, future research should examine whether some types of questions generate larger learning gains than others.

Another perspective for further research would be to assess whether tests are required to produce higher learning or whether other practices involving retrieval could lead to similar positive effects. For example, the literature suggests that judgment of learning (JOL) could be a good predictor of subsequent memory performance particularly for relatively simple material, especially when delayed (Dunlosky & Nelson, 1992). JOL tasks only consist of asking participants to report how able they think they will be to retrieve the information on a subsequent test, and such a task could be sufficient to induce retrieval (Nelson & Dunlosky, 1991). If the reason why tests increase performance is because they involve retrieval effort, any practice that involves retrieval (including JOL) should be efficient in increasing learning. Supporting this hypothesis, laboratory studies conducted with pairs of words showed that JOL was just as efficient as tests for improving subsequent memory performance (Akdoğan, Izaute, Danion, Vidailhet, & Bacon, 2016; Jönsson, Hedner, & Olsson, 2012). However, a meta-analysis of literature on this phenomenon recently highlighted that the type of material seems to drives it (Double, Birney, & Walker, 2018). Indeed, the enhancing effect of JOLs on learning, called “reactivity effect”, is moderate in the case of related word pairs but not significant for unrelated or a mixture of unrelated and related pairs. Similarly, although not conducted in a real classroom context, in a recent set of five experiments, Ariel et al. (2020), tested the effect of these practices on the retention of material more alike as educational relevant material than word pairs. Their results showed that JOLs in their standard form did not enhance learning of this material whereas JOLs with retrieval instructions did. Future research should test whether JOL, especially with retrieval instructions, used in a real classroom context, could also increase academic performance.

Another important criteria to take into account is the delay between the last test and the exam. Indeed, to observe the beneficial effect of test, a delay between the retrieval practice and the exam is necessary. The reviewed studies here that showed a testing effect reported a delay with the exception of Greving, Lenhard & Richter's (2020) that only obtained such an effect for delay longer than 21 days, the others obtained it with more shorter delays.

Testing for learning, not for ranking

Ultimately, the findings of the reviewed studies give very important insights into how tests should be implemented in the classroom, but they also point to important aspects that should be taken into account when considering tests in real classroom contexts. As discussed herein, the classroom context is not a neutral context, but an evaluative context (Baumeister, 1984; Huguet & Kuyper, 2008) that is likely to elicit various forms of motivations and goals. According to social psychology research, tests are one characteristic likely to increase performance goals, social comparison concerns, and evaluative pressure (Ames, 1992; Darnon et al., 2012; Meece et al., 2006; Pulfrey, Buchs, & Butera, 2011). In such a context, tests are particularly likely to elicit anxiety and threat with dramatic consequences for learning (Cassady, 2004), particularly among vulnerable students (Croizet, Goudeau, Marot, & Millet, 2017; Jury et al., 2015). Supporting this idea, the research reviewed in the present paper points to the fact that tests are more likely to increase learning if they occur in a collaborative (versus competitive) context (Cranney et al., 2009; Eastridge & Benson, 2020; Vojdanova et al., 2010; Wiklund-Hörnqvist et al., 2014) or when they are non-graded (versus graded; Khanna, 2015) or with only a low impact on grade (Johnson & Kiviniemi, 2009). This is also strongly supported by a recent meta-analysis dedicated to this question (Sotola & Crede, 2020) showing a positive association between the uses of in-class low-stakes quizzes on later exam performance and on the odds of passing a class. But, in all cases, the fact that the test is

compulsory or as least strongly encouraged (Grimstad & Grabe, 2004; Trumbo et al., 2016) affects the occurrence of such a beneficial effect of testing on learning. Such conclusions are totally in line with social psychology research showing that tests can be perceived as either learning tools or selection tools with further impact on learning outcomes (Darnon et al., 2011; Smeding et al., 2013).

Related to this point, the question of whether feedback is required for the test effect to occur is of great interest for teachers. Indeed, providing formative feedback is an important part of usual teaching practices as it allows educators to determine what is already understood and known and what is not (Black & Wiliam, 2003). In the studies reviewed herein, it is difficult to draw clear conclusions on the effect of feedback because none of the reviewed studies directly compared testing effect with and without feedback. Moreover, some of the studies offered participants corrective feedback (e.g. E. L. Bjork et al., 2014; Carpenter et al., 2018; Khanna, 2015; Thomas et al., 2020; Welch, 2019), including one with delayed feedback (E. L. Bjork et al., 2014), but all studies found a testing effect, even when no feedback was provided (Greving & Richter, 2018). Thus, evidence is lacking in classroom settings to draw definitive conclusions on this issue. Yet one meta-analysis conducted with real classroom experiments (Schwieren et al., 2017) showed that, when other moderators were held constant in the analysis, the feedback enhanced the positive effect of test on learning, although the evidence seems less consistent in a more recent literature review (Moreira et al., 2019). Similarly, although based on lab studies, Rowland (2014) highlighted that studies where tests were associated with feedback produced greater effect than tests without feedback. This conclusion is consistent with what was obtained in the field of learning language. Indeed, many studies in this field have shown that feedback improved the positive effect of testing on learning (e.g., Hays, Kornell, & Bjork, 2010; Metcalfe, Kornell, & Finn, 2009; Pashler, Cepeda, Wixted, & Rohrer, 2005), particularly when the feedback included the correct answer

(Pashler et al., 2005). Butler et al. (2008) further documented that delayed feedback would lead to greater benefits than immediate feedback. To summarize, even if evidence is still lacking in classroom contexts, lab studies suggest that providing corrective feedback should increase the positive effect of testing on learning. This idea is consistent with the fact that, as developed herein, tests are more likely to produce a learning gain when perceived as learning and formative tools (rather than selection tools). Indeed, formative feedback is one pedagogical practice likely to increase mastery and learning goals (Meece, Anderman, & Anderman, 2006). We believe the moderating role of feedback on the testing effect in real classroom contexts warrants further exploration and should be the subject of future research in the field. First of all, some efforts to clarify what a feedback is in this context appear to be necessary in order to, eventually, determine how it varies, then whether or not it influences the testing effect. In particular, as discussed above, there are inconsistencies on the best question format (SAQ or MCQ) to obtain the more efficient testing. It seems, in particular, that studies which document a beneficial effects of MCQ are all classroom studies (Moreira et al., 2019) and all includes feedbacks. Thus, it seems reasonable to argue that MCQ without feedback would not raise similar positive effects and that the actual beneficial effect of MCQ in classrooms could be indirectly due to the positive effects of feedback. Testing whether the presence of feedbacks moderates the effect of question format would in that sense represent an interesting question to investigate in future research.

Finally, students' socioeconomic status (SES) is a variable that has rarely been examined in the test literature but might affect students' reactions to tests and, consequent, affect performance, particularly in real classroom contexts. Indeed, abundant literature in social psychology has shown that, in real classroom contexts, a SES achievement gap appears; the higher the evaluative pressure, the greater the gap (e.g., Croizet & Claire, 1998; Goudeau & Croizet, 2017). Thus, on the one hand, in real classroom contexts, tests could

increase evaluative pressure and thereby increase the probability that the SES achievement gap will occur. If this reasoning is true, the test should have more positive effects on higher SES students than on lower SES students. On the other hand, as discussed herein, when used as learning tools, tests can decrease the SES achievement gap (Pennebaker, Gosling, & Ferrell, 2013; Smeding, Darnon, Souchal, Toczek-Capelle, & Butera, 2013). Further research should explore these two alternative hypotheses in order to determine whether SES moderates the testing effects in real classroom contexts or not and, if so, in which direction.

Conclusion

Taken together, the results of the studies reviewed in the present paper support the existence of a testing effect in the ecological context of a classroom. This review also highlights important features regarding the implementation of tests in classrooms that can prove particularly useful for teachers. For example, depending on the time teachers can allocate to testing students during class time, they can use either multiple choice or short answer questions. If teachers prefer not to allocate class time to testing, an option could be to implement their tests outside of the classroom, such as with e-learning devices. Finally, although more research is needed to draw definitive conclusions, we encourage teachers to include corrective or informative feedback after the test questions, particularly because corrective feedback should increase the formative value and meaning of tests and, consequently, increase the likelihood that students will perceive them as learning tools rather than selection tools, with further benefits for learning.

As mentioned in the introduction of the present paper, despite the wide literature exploring the testing effect, not many studies have focused on its actual benefit for students' general academic achievement. Authors often conclude that tests are an efficient learning tool that would benefit students by being implemented in classrooms, yet quite few studies have, thus far, examined whether tests—more than restudy or other non-retrieval practice—really

increase learning in a real class. As discussed herein, we think this is a very important limitation because there are reasons to believe that the very meaning of a test might be quite different in lab settings and in real classroom settings. In the present article, we have reviewed the studies that did examine the benefits of tests in class at university, we found an overall positive effect. Replications are required in class to strengthen these promising results and delimitate the conditions for such a positive effect to occur in real classroom contexts.

References

Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the Use of Tests: A Meta-Analysis of Practice Testing. *Review of Educational Research*, *87*(3), 659–701. https://doi.org/10.3102/0034654316689306

Akdoğan, E., Izaute, M., Danion, J. M., Vidailhet, P., & Bacon, E. (2016). Is retrieval the key? Metamemory judgment and testing as learning strategies. *Memory*, *24*(10), 1390–1395. https://doi.org/10.1080/09658211.2015.1112812

Ariel, R., Karpicke, J. D., Witherby, A. E., & Tauber, S. K. (2020). Do Judgments of Learning Directly Enhance Learning of Educational Materials? *Educational Psychology Review*. https://doi.org/10.1007/s10648-020-09556-8

Arnold, K. M., & Mcderm.ott, K. B. (2013). Test-potentiated Learning: Distinguishing between direct and indirect effects of tests. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *39*(3), 1–14. https://doi.org/10.1037/a0029199.Test-Potentiated

Balch, W. R. (1998). Practice versus Review Exams and Final Exam Performance. *Teaching of Psychology*, *25*(3), 181–185. https://doi.org/10.1207/s15328023top2503_3

Batsell, W. R., Perry, J. L., Hanley, E., & Hostetter, A. B. (2017). Ecological Validity of the Testing Effect: The Use of Daily Quizzes in Introductory Psychology. *Teaching of Psychology*, *44*(1), 18–23. https://doi.org/10.1177/0098628316677492

Bell, M. C., Simone, P. M., & Whitfield, L. C. (2015). Failure of online quizzing to improve performance in introductory psychology courses. *Scholarship of Teaching and Learning in Psychology*, *1*(2), 163–171. https://doi.org/10.1037/stl0000020

Bjork, E. L., Little, J. L., & Storm, B. C. (2014). Multiple-Choice Testing as a Desirable Difficulty in the Classroom. *Journal of Applied Research in Memory and Cognition*, *3*(3), 165–170. https://doi.org/10.1016/j.jarmac.2014.03.002

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-Regulated Learning : Beliefs , Techniques , and Illusions. *Annual Review of Psychology*, *64*, 417–444. https://doi.org/10.1146/annurev-psych-113011-143823

Black, P., & Wiliam, D. (2003). "In Praise of Educational Research": Formative assessment. *British Educational Research Journal*, *29*(5), 623–637. https://doi.org/10.1080/0141192032000133721

Butler, A. C., Karpicke, J. D., & Roediger, H. L. (2008). Correcting a metacognitive error: feedback increases retention of low-confidence correct responses. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *34*(4), 918–928. https://doi.org/10.1037/0278-7393.34.4.918

Butler, A. C., Marsh, E. J., Slavinsky, J. P., & Baraniuk, R. G. (2014). Integrating Cognitive Science and Technology Improves Learning in a STEM Classroom. *Educational Psychology Review*, *26*(2), 331–340. https://doi.org/10.1007/s10648-014-9256-4

Butler, A. C., & Roediger, H. L. (2007). Testing improves long-term retention in a simulated classroom setting. *European Journal of Cognitive Psychology*, *19*(4–5), 514–527. https://doi.org/10.1080/09541440701326097

Carpenter, S. K. (2012). Testing Enhances the Transfer of Learning. *Current Directions in Psychological Science*, *21*(5), 279–283. https://doi.org/10.1177/0963721412452728

Carpenter, S. K., & Delosh, E. L. (2005). Application of the testing and spacing effects to name learning. *Applied Cognitive Psychology*, *19*(5), 619–636. https://doi.org/10.1002/acp.1101

Carpenter, S. K., & Delosh, E. L. (2006). Impoverished cue support enhances subsequent retention : Support for the elabo ... *Memory*, (2), 268–276.

Carpenter, S. K., Rahman, S., & Perkins, K. (2018). The effects of prequestions on classroom learning. *Journal of Experimental Psychology: Applied*, *24*(1), 34–42.

https://doi.org/10.1037/xap0000145

Cepeda, N., Pashler, H., Vul, E., Wixted, J., & Rohrer, D. (2006). Distributed practice in verbal recall tasks : A review and quantitative synthesis. *Psychological Bulletin*, *132*(3), 354–380. Retrieved from https://cloudfront.escholarship.org/dist/prd/content/qt3rr6q10c/qt3rr6q10c.pdf

Chan, J. C. K., Meissner, C. A., & Davis, S. D. (2018). Retrieval potentiates new learning: A theoretical and meta-analytic review. *Psychological Bulletin*, *144*(11), 1111–1146. https://doi.org/10.1037/bul0000166

Chang, S. H. (2018). Testing effect in a college class. *Psychology and Education: An Interdisciplinary Journal*, *55*(1–2), 41–46.

Connolly, P., Keenan, C., & Urbanska, K. (2018). The trials of evidence-based practice in education: a systematic review of randomised controlled trials in education research 1980–2016. *Educational Research*, *60*(3), 276–291. https://doi.org/10.1080/00131881.2018.1493353

Cranney, J., Ahn, M., McKinnon, R., Morris, S., & Watts, K. (2009). The testing effect, collaborative learning, and retrieval-induced facilitation in a classroom setting. *European Journal of Cognitive Psychology*, *21*(6), 919–940. https://doi.org/10.1080/09541440802413505

Croizet, J.-C., & Claire, T. (1998). Extending the concept of stereotype threat to social class: The intellectual underperformance of students from low socioeconomic backgrounds. *Personality and Social Psychology Bulletin*, *24*(6), 588–594. https://doi.org/10.1177/0146167298246003

Darnon, C., Smeding, A., Toczek-Capelle, M.-C., & Souchal, C. (2011). L'évaluation comme outil de formation et/ou de sélection. In F. Butera, C. Buchs, & C. Darnon (Eds.), *L'évaluation, une menace ?* (pp. 114–125). Paris: PUF.

Dehaene, S. (2019). *Les Sciences au Service de l'Ecole*. Paris: Odile Jacob Réseau Canopé.

Dobson, J. L., & Linderholm, T. (2014). Self-testing promotes superior retention of anatomy and physiology information. *Advances in Health Sciences Education*, *20*(1), 149–161. https://doi.org/10.1007/s10459-014-9514-8

Double, K. S., Birney, D. P., & Walker, S. A. (2018). A meta-analysis and systematic review of reactivity to judgements of learning. *Memory*, *26*(6), 741–750. https://doi.org/10.1080/09658211.2017.1404111

Dunlosky, J., & Nelson, T. O. (1992). Importance of the kind of cue for judgments of learning (JOL) and the delayed-JOL effect. *Memory & Cognition*, *20*(4), 374–380. https://doi.org/10.3758/BF03210921

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest, Supplement*, *14*(1), 4–58. https://doi.org/10.1177/1529100612453266

Eastridge, J. A., & Benson, W. L. (2020). Comparing Two Models of Collaborative Testing for Teaching Statistics. *Teaching of Psychology*, *47*(1), 68–73. https://doi.org/10.1177/0098628319888113

Einstein, G. O., Mullet, H. G., & Harrison, T. L. (2012). The Testing Effect: Illustrating a Fundamental Concept and Changing Study Strategies. *Teaching of Psychology*, *39*(3), 190–193. https://doi.org/10.1177/0098628312450432

Ferguson, C. J., & Brannick, M. T. (2012). Publication bias in psychological science: Prevalence, methods for identifying and controlling, and implications for the use of meta-analyses. *Psychological Methods*, *17*(1), 120–128. https://doi.org/10.1037/a0024445

Foss, D. J., & Pirozzolo, J. W. (2017). Four Semesters Investigating Frequency of Testing ,

the Testing Effect , and Transfer of Training. *Journal of Educational Psychology*, *109*(8), 1067–1083. https://doi.org/10.1037/edu0000197

Francis, G. (2012). Publication bias and the failure of replication in experimental psychology. *Psychonomic Bulletin and Review*, *19*(6), 975–991. https://doi.org/10.3758/s13423-012-0322-y

Goudeau, S., & Croizet, J.-C. (2017). Hidden Advantages and Disadvantages of Social Class: How Classroom Settings Reproduce Social Inequality by Staging Unfair Comparison. *Psychological Science*, *28*(2), 162–170. https://doi.org/10.1177/0956797616676600

Greving, S., Lenhard, W., & Richter, T. (2020). Adaptive retrieval practice with multiple-choice questions in the university classroom. *Journal of Computer Assisted Learning*, 1–11. https://doi.org/10.1111/jcal.12445

Greving, S., & Richter, T. (2018). Examining the Testing Effect in University Teaching : Retrievability and Question Format Matter. *Frontiers in Psychology*, 1–10. https://doi.org/10.3389/fpsyg.2018.02412

Grimstad, K., & Grabe, M. (2004). Are Online Study Questions Beneficial? *Teaching of Psychology*, *31*(2), 143–146. https://doi.org/10.1207/s15328023top3102_8

Hays, M. J., Kornell, N., & Bjork, R. A. (2010). The costs and benefits of providing feedback during learning. *Psychonomic Bulletin and Review*, *17*(6), 797–801. https://doi.org/10.3758/PBR.17.6.797

Hays, M. J., Kornell, N., & Bjork, R. A. (2013). When and why a failed test potentiates the effectiveness of subsequent study. *Journal of Experimental Psychology: Learning Memory and Cognition*, *39*(1), 290–296. https://doi.org/10.1037/a0028468

Inouye, C. Y., Bae, C. L., & Hayes, K. N. (2017). Using whiteboards to support college students ' learning of complex physiological concepts. *Advances in Physiological Education*, *41*, 478–484. https://doi.org/10.1152/advan.00202.2016

Johnson, B. C., & Kiviniemi, M. T. (2009). The Effect of Online Chapter Quizzes on Exam Performance in an Undergraduate Social Psychology Course. *Teaching of Psychology*, *36*(1), 33–37. https://doi.org/10.1080/00986280802528972

Jönsson, F. U., Hedner, M., & Olsson, M. J. (2012). The Testing Effect as a Function of Explicit Testing Instructions and Judgments of Learning, *59*(5), 251–257. https://doi.org/10.1027/1618-3169/a000150

Jury, M., Smeding, A., & Darnon, C. (2015). First-generation students' underperformance at university: the impact of the function of selection. *Frontiers in Psychology*, *6*(May), 710. https://doi.org/10.3389/fpsyg.2015.00710

Kang, S. H. K., & Pashler, H. (2014). Is the benefit of retrieval practice modulated by motivation ? *Journal of Applied Research in Memory and Cognition*, *3*, 183–188. https://doi.org/10.1016/j.jarmac.2014.05.006

Karpicke, J. D. (2017). Retrieval-based learning: A decade of progress. In J. H. Byrne (Ed.), *Learning and memory: A comprehensive reference* (2nd ed., Vol. 1–4, pp. 487–514). Elsevier. https://doi.org/10.1016/B978-0-12-809324-5.21055-9

Karpicke, J., & Roediger, H. (2005). Test-Enhanced Learning: Taking Memory Tests Improves Long-Term Retention. *Psychological Science*, *17*(3), 249–255.

Kelley, M. R., Chapman-Orr, E. K., Calkins, S., & Lemke, R. J. (2019). Generation and Retrieval Practice Effects in the Classroom Using PeerWise. *Teaching of Psych*, *46*(2), 121–126. https://doi.org/10.1177/0098628319834174

Khanna, M. M. (2015). Ungraded Pop Quizzes : Test-Enhanced Learning Without All the Anxiety. *Teaching of Psychology*, *42*(2), 174–178. https://doi.org/10.1177/0098628315573144

Kift, S. (2015). A decade of Transition Pedagogy : A quantum leap in conceptualising the first year experience. *HERDSA Review of Higher Education, 12*, 51–86.

Kornell, N., Hays, M. J., & Bjork, R. A. (2009). Unsuccessful Retrieval Attempts Enhance Subsequent Learning. *Journal of Experimental Psychology: Learning Memory and Cognition*, *35*(4), 989–998. https://doi.org/10.1037/a0015729

Kornell, N., Rabelo, V. C., & Jacobs Klein, P. (2012). Tests enhance learning — Compared to what ? *Journal of Applied Research in Memory and Cognition*, *1*, 257–259. https://doi.org/10.1016/j.jarmac.2012.10.002

Kornell, N., & Son, L. K. (2009). Learners' choices and beliefs about self-testing. *Memory*, *17*(5), 493–501. https://doi.org/10.1080/09658210902832915

Leeming, F. C. (2002). The Exam-A-Day Procedure Improves Performance in Psychology Classes. *Teaching of Psychology*, *29*(3), 210–212. https://doi.org/10.1207/S15328023TOP2903_06

Little, J. L., Bjork, E. L., Bjork, R. A., & Angello, G. (2012). Multiple-Choice Tests Exonerated, at Least of Some Charges: Fostering Test-Induced Learning and Avoiding Test-Induced Forgetting. *Psychological Science*, *23*(11), 1337–1344. https://doi.org/10.1177/0956797612443370

Lyle, K. B., & Crawford, N. A. (2011). Retrieving Essential Material at the End of Lectures Improves Performance on Statistics Exams. *Teaching of Psychology*, *38*(2), 94–97. https://doi.org/10.1177/0098628311401587

McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology*, *19*(4–5), 494–513. https://doi.org/10.1080/09541440701326154

Metcalfe, J., Kornell, N., & Finn, B. (2009). Delayed versus immediate feedback in children's and adults' vocabulary learning. *Memory and Cognition*, *37*(8), 1077–1087. https://doi.org/10.3758/MC.37.8.1077

Moreira, B. F. T., Pinto, T. S. S., Starling, D. S. V., & Jaeger, A. (2019). Retrieval Practice in

Classroom Settings: A Review of Applied Research. *Frontiers in Education*, *4*(February). https://doi.org/10.3389/feduc.2019.00005

Ndao, G., & Ministère de l'Enseignement Supérieur de la Recherche et de l'Innovation. (2019). Les étudiants en formation dans l'enseignement supérieur. Retrieved January 10, 2020, from https://publication.enseignementsup-recherche.gouv.fr/eesr/FR/T191/les_etudiants_en_formation_dans_l_enseignement_superieur/

Nelson, T. O., & Dunlosky, J. (1991). When People's Judgments of Learning (JOLs) are Extremely Accurate at Predicting Subsequent Recall: The "delayed-JOL effect." *Psychological Science*, *2*(4), 267–270. https://doi.org/10.1111/j.1467-9280.1991.tb00147.x

Nungester, R. J., & Duchastel, P. C. (1982). Testing versus review: Effects on retention. *Journal of Educational Psychology*, *74*(1), 18–22. https://doi.org/10.1037/0022-0663.74.1.18

Pan, S. C., & Rickard, T. C. (2018). Transfer of Test-Enhanced Learning: Meta-Analytic Review and Synthesis. *Psychological Bulletin*, *144*(7), 710–756. https://doi.org/10.1037/bul0000151

Pashler, H., Cepeda, N. J., Wixted, J. T., & Rohrer, D. (2005). When does feedback facilitate learning of words? *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *31*(1), 3–8. https://doi.org/10.1111/j.1365-2648.2008.04878.x/abstract

Pashler, H., Rohrer, D., Cepeda, N. J., & Carpenter, S. K. (2007). Enhancing learning and retarding forgetting: Choices and consequences. *Psychonomic Bulletin and Review*, *14*(2), 187–193. https://doi.org/10.3758/BF03194050

Pennebaker, J. W., Gosling, S. D., & Ferrell, J. D. (2013). Daily online testing in large classes: Boosting college performance while reducing achievement gaps. *PLoS ONE*,

8(11). https://doi.org/10.1371/journal.pone.0079774

Pyc, M. A., & Rawson, K. A. (2007). Examining the efficiency of schedules of distributed retrieval practice. *Memory and Cognition*, *35*(8), 1917–1927. https://doi.org/10.3758/BF03192925

Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language*, *60*(4), 437–447. https://doi.org/10.1016/j.jml.2009.01.004

Pyc, M. A., & Rawson, K. A. (2012). Why is test-restudy practice beneficial for memory? An evaluation of the mediator shift hypothesis. *Journal of Experimental Psychology: Learning Memory and Cognition*, *38*(3), 737–746. https://doi.org/10.1037/a0026166

Roediger, H. L., & Karpicke, J. D. (2006a). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*(3), 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

Roediger, H. L., & Karpicke, J. D. (2006b). The power of testing memory: basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*(3), 181–210.

Roediger, H. L., Putnam, A. L., & Smith, M. A. (2011). Ten Benefits of Testing and Their Applications to Educational Practice. In *Psychology of Learning and Motivation - Advances in Research and Theory* (Vol. 55, pp. 1–36). https://doi.org/10.1016/B978-0-12-387691-1.00001-6

Rowland, C. A. (2014). The Effect of Testing Versus Restudy on Retention : A Meta-Analytic Review of the Testing Effect. *Psychological Bulletin*, *140*(6), 1432.

Schwieren, J., Barenberg, J., & Dutke, S. (2017). The Testing Effect in the Psychology Classroom : A Meta-Analytic Perspective. *Psychology Learning & Teaching*, *16*(2), 179–196. https://doi.org/10.1177/1475725717695149

Shapiro, A. M., & Gordon, L. T. (2012). A Controlled Study of Clicker-Assisted Memory Enhancement in College Classrooms. *Applied Cognitive Psychology*, *26*(4), 635–643. https://doi.org/10.1002/acp.2843

Smeding, A., Darnon, C., Souchal, C., Toczek-Capelle, M.-C., & Butera, F. (2013). Reducing the Socio-Economic Status Achievement Gap at University by Promoting Mastery-Oriented Assessment. *PLoS ONE*, *8*(8), 1–6. https://doi.org/10.1371/journal.pone.0071678

Sotola, L. K., & Crede, M. (2020). Regarding Class Quizzes: a Meta-analytic Synthesis of Studies on the Relationship Between Frequent Low-Stakes Testing and Class Performance. *Educational Psychology Review*. https://doi.org/10.1007/s10648-020-09563-9

Stephens, N. M., Markus, H. R., & Fryberg, S. A. (2012). Social class disparities in health and education: Reducing inequality by applying a sociocultural self model of behavior. *Psychological Review*, *119*(4), 723–744. https://doi.org/10.1037/a0029028

Thomas, A. K., Smith, A. M., Kamal, K., & Gordon, L. T. (2020). Should You Use Frequent Quizzing in Your College Course? Giving up 20 Minutes of Lecture Time May Pay Off. *Journal of Applied Research in Memory and Cognition*, *9*(1), 83–95. https://doi.org/10.1016/j.jarmac.2019.12.005

Trumbo, M. C., Leiting, K. A., McDaniel, M. A., & Hodge, G. K. (2016). Effects of Reinforcement on Test-Enhanced Learning in a Large, Diverse Introductory College Psychology Course. *Journal of Experimental Psychology: Applied*, *22*(2), 148–160. https://doi.org/10.1037/xap0000082

Vojdanova, M., Cranney, J., & Newell, B. R. (2010). The Testing Effect: The Role of Feedback and Collaboration in a Tertiary Classroom Setting. *Applied Cognitive Psychology*, *24*(24), 1183–1195. https://doi.org/10.1002/acp

Welch, S. (2019). Scholarship of MacMillan Education's LaunchPas as a Textbook Technology Supplement When Teaching Introductory Psychlogy. *Scolarship of Teaching and Learning in Psychology*, *5*(3), 236–247. https://doi.org/10.1037/stl0000162

Wheeler, M. A., & Roediger, H. L. (1992). Disparate effects of repeated testing: Reconciling Ballard's (1913) and Bartlett's (1932) results. *Psychological Science*, *3*(4), 240–245.

Wiklund-Hörnqvist, C., Jonsson, B., & Nyberg, L. (2014). Strengthening concept learning by repeated testing. *Scandinavian Journal of Psychology*, *55*(1), 10–16. https://doi.org/10.1111/sjop.12093

Wissman, K. T., & Rawson, K. A. (2018). Test-potentiated learning: three independent replications, a disconfirmed hypothesis, and an unexpected boundary condition. *Memory*, *26*(4), 406–414. https://doi.org/10.1080/09658211.2017.1350717

Zaromb, F. M., & Roediger, H. L. (2010). The testing effect in free recall is associated with enhanced organizational processes. *Memory and Cognition*, *38*(8), 995–1008. https://doi.org/10.3758/MC.38.8.995