Desirable Difficulties in Applied Learning Settings: Mechanisms and Effects

Inaugural-Dissertation

zur Erlangung der Doktorwürde der Philosophie (Dr. phil.)

Fakultät für Humanwissenschaften

der

Julius-Maximilians-Universität Würzburg

Vorgelegt von

Sven Greving

aus Kassel

Kassel, Januar 2021

Betreuer: Prof. Dr. Tobias Richter

Erstgutachter: Prof. Dr. Wolfgang Lenhard

Zweitgutachter: Prof. Dr. Ralf Rummer (Universität Kassel)

Tag der Disputation: 29.10.2021

Summary

Improving retention of learned content by means of a practice test is a learning strategy that has been researched since a century and has been consistently found to be more effective than comparable learning strategies such as restudy (i.e., the *testing effect*). Most importantly, practicing test questions has been found to outperform restudy even when no additional information about the correct answers was provided to practice test takers, rendering practice tests effective and efficient in fostering retention of learning content. Since 15 years, additional scientific attention is devoted to this memory phenomenon and additional research investigated to what extend practicing test questions is relevant in real-world educational settings. This dissertation first presents the evidence for testing effects in applied educational settings by presenting key publications and presenting findings from a methodological review conducted for this purpose. Within this dissertation, theories are presented why practicing test questions should benefit learning in real-world educational settings even without the provision of additional information and key variables for the effectiveness of practicing test questions are presented. Four studies presented in this dissertation aimed at exploring these assumptions in actual university classrooms while also trying to implement new methods of practicing learning content and thus augment course procedures. Findings from these studies—although not often consistent—will be incorporated and interpreted in the light of the theoretical accounts on the testing effect. The main conclusion that can be drawn from this dissertation is that, given the right circumstances, practicing test questions can elicit beneficial effects on the retention of learning content that are independent of additional information and thus taking a practice test per se, can foster retention of real-world learning content.

Zusammenfassung

Seit einem Jahrhundert wird die positive Wirkung von Abrufübungen auf das Behalten gelernter Inhalte untersucht. Diese Untersuchungen belegen durchgängig, dass Abrufübungen für das Behalten förderlicher sind als vergleichbare Lernaktivitäten wie beispielsweise nochmaliges Lesen (Der *Testungseffekt*). Der Testungseffekt ist besonders bedeutsam wenn er in Settings untersucht wird in denen keine Informationen über die Richtigkeit des abgelegten Tests an die Lernenden gegeben werden und somit belegen, dass Abrufübungen nicht nur effektive sondern auch effiziente Methoden der Behaltensförderung darstellen. Seit 15 Jahren intensiviert sich das wissenschaftliche Interesse am Testungseffekt und immer mehr Forschung untersucht auch das Ausmaß der Übertragbarkeit auf reale Lernumgebungen. Diese Dissertation stellt zuerst Befunde aus Untersuchungen des Testungseffekts in angewandten Lernsettings dar, wobei zentrale Studien besprochen werden und hier ein eigens zu diesem Zweck angestelltes methodisches Review vorgestellt wird. Innerhalb der Dissertationen werden Theorien erläutert, warum Abrufübungen in realen Lernsettings das Behalten auch ohne zusätzliche Darbietung der korrekten Antwort fördern sollten und welche Variablen dafür essentiell sind. In dieser Dissertation warden zudem vier Studien präsentiert, die diese Variablen in universitären Lernumgebungen untersuchten, teilweise mit dem Ziel, Abrufübungen noch effektiver zu machen. Die gewonnenen Erkenntnisse aus diesen vier Studien—obwohl sie nicht immer konsistent sind—werden abschließend diskutiert und in Bezug zu den vorgestellten Theorien gesetzt. Eine zetrale Schlussfolgerung aus den vorgestellten Studien ist die Erkenntnis, dass unter den richtigen Bedingungen Abrufübungen den Testungseffekt hervorrufen können, die unabhängig von zusätlicher Information sind und dass es demnach die Abrufübung per se ist, die zum Behalten gelernter Information beiträgt.

Table of Contents

Introduction

Tests are ubiquitous in university teaching, mainly as means of summative assessment. Although often feared and loathed by learners (Khanna, 2015), they are not only a tool for the assessment of learning but also for improving learning. Practice tests that require learners to answer questions about learning content are also often used by students and teachers to foster retention of learned content for example at the end of a lecture or after reading a book chapter. Indeed, researchers rank practicing by testing as one of the most effective ways to increase retention (Dunlosky et al., 2013) and consequently propagate the use of this practice strategy (Dunlosky & Rawson, 2015; Dunn et al., 2013). This recommendation is backed by nearly a century of research investigating the effects of testing (for a review of that time span, see Phelps, 2012): Subsequent to first studying learning material, testing has been shown to be more effective in terms of retention than other strategies like additional restudying (e.g., Roediger & Karpicke, 2006a), note-taking (Rummer et al., 2017), closed-book tests (e.g. Rummer et al., 2019), group discussions (Stenlund et al., 2017), or even practice of acquired skills (Kromann et al., 2009). This superiority of testing is called the testing effect, the retrieval practice effect, or retrieval-induced learning.

When recommending retrieval practice in an educational setting, researchers assume that taking a test presents a difficulty to learners that has to be overcome. For this reason, practice testing is considered a desirable difficulty, which is a learning strategy that makes learning more difficult and seemingly less effective in the short run but rewards learners with increased retention for learning material that has been practiced this way (R. A. Bjork, 1994). Other desirable difficulties that follow this rational are the benefits that arise from longer time intervals between repeated learning activities (i.e., the spacing effect), alternating between to-

be learned subjects in order to create difficulties by mentally differentiating between the two (i.e., the effect of interleafing) and generating examples, questions or mind-maps instead of restudying (i.e., the generation effect).

Whenever researchers recommend practicing tests they expect that this difficulty (i.e., the test) results in two types of beneficial effects on retention (Dunlosky et al., 2013; Dunlosky & Rawson, 2015; Roediger & Karpicke, 2006b). Testing can lead to direct effects on memory, which means that the process of testing oneself, or more precisely by retrieving the information from one's memory improves, and so accessing this information becomes easier and thus promoting retention of the retrieved information. Testing can also lead to indirect effects by triggering processes that promote retention like demonstrating knowledge gaps to test takers or motivating students to study in expectancy of a test. This distinction between direct and indirect testing effects will be discussed more in detail later.

Findings from surveys among university students indicate that learners are not blind to both direct and indirect testing effects. Several studies among university students indicate that the majority of students (80%–86%) regularly tests themselves, however only few students (18%–31%) think that practice alone—without feedback and additional restudy—is beneficial for retention (Hartwig & Dunlosky, 2012; Kornell & Bjork, 2007; Morehead et al., 2016). Further evidence from studies investigating simulated learning environments indicate a similar pattern when forcing learners to either restudy or test themselves after initial learning: Most students chose to test themselves but only few saw testing itself as beneficial for retention (Karpicke et al., 2009; Kornell & Son, 2009).

The main goal of this dissertation is to investigate the beneficial effects of practice tests on retention in real-world educational settings, such as university courses. To do so, this dissertation aims to advance the understanding in three regards.

The first aim concerns the investigation of direct testing effects in applied educational settings. Contrary to the assumptions of most students, proponents of the testing effect expect beneficial outcomes of taking practice tests even when test takers do not receive feedback in the form of presenting them the correct answers. Data from laboratory research suggests that learning from testing alone is beneficial (e.g., Kornell & Son, 2009; Roediger & Karpicke, 2006b), however this claim is hardly backed up by data from real-world educational settings (e.g., Moreira et al., 2019). It seems reasonable to assume that taking a test and then receiving feedback on the given answers is beneficial for learning because it triggers indirect testing effects such as metacognitive monitoring, additional studying, and an additional presentation of the correct answer (McDaniel & Little, 2019). But is it also reasonable to assume, that practice tests are still superior to restudying and that this difference is relevant in real-world educational contexts, when controlling for factors that elicit indirect testing effects (i.e., withholding feedback)?

The second aim of this dissertation is to examine the effects of test question difficulty as a relevant factor that moderates the testing effect in real-world educational contexts. Initially it has been stated that the beneficial effects of practice tests are thought to be a result of desirable difficulties learners face when practicing tests. It should be noted that in real-world educational contexts, answering questions about learning content varies in difficulty within the same practice test, between learners with differing characteristics. Furthermore, research indicates that learners have different perceptions of difficulty for diverse question formats: Multiple-choice tests seem to be favored over other test types because learners associate multiple-choice questions with lesser difficulty and more expected success (Struyven et al., 2005; Zeidner, 1987). First evidence suggests that lower difficulty might also render practice questions more attractive than other learning strategies: When

difficulty was manipulated andquestions made easier by providing more cues to the correct answer, participants preferred easier questions to rereading (Toppino et al., 2018; Vaughn & Kornell, 2019).

Laboratory research indicates that the difficulty of test questions in general as well as cues to the correct answers in particular mediate the testing effect (e.g., Finley et al., 2011; Rowland, 2014). In the research of testing effects, difficulty has been conceptualized in various ways: Some studies varied context factors that are assumed to have an effect on the accessibility of practiced learning content (e.g., Carpenter, 2009; Pyc & Rawson, 2009) while others varied the amount of helpful cues in the practice tests (Finley et al., 2011). Furthermore, practice tests in real-world educational settings might differ in their question format. Different question formats have been associated with different difficulty levels (Glover, 1989). Additionally, practice test question difficulty is linked to the performance in the practice tests and thus the amount of information correctly retrieved from memory (i.e., retrievability): With higher difficulty of practice questions, the success of answering these questions decreases and vice versa (Kornell & Vaughn, 2016). On the basis of this link it is possible to infer practice test difficulty from the inverse retrievability.

Meta-analytic evidence from laboratory studies indicates that retrievability is linked to more pronounced direct testing effects (Rowland, 2014). However, higher difficulty might also result in more effort spent to overcome this difficulty and thus ultimately result in better retention (R. A. Bjork, 1994).

The third aim of this dissertation is to capitalize on the knowledge about the role difficulty plays in direct testing effects in real-world educational contexts and investigate whether altering the difficulty of practice tests bears the potential of capitalizing on the testing effect in actual educational settings.

To achieve these three aims, four empirical studies have been conducted. The first empirical study (Chapter II) presented here investigates the question whether different question formats can elicit direct testing effects in an actual university lecture. It furthermore tries to shed light on the question whether retrievability of learning content as a product of differing question difficulties moderates direct testing effects. The second empirical study (Chapter III)—a conceptual replication of the first study— further explores the extent to which question format and retrievability are responsible for direct testing effects observed in real-world educational contexts. The remaining empirical studies investigate means to enhance the effectiveness of practicing multiple-choice questions. In the third empirical study (Chapter IV), the role of retrievability and feedback are further explored in a field experiment, whereas feedback was intended to present an additional difficulty because learners had to revisit specific parts of the learned content, resulting in additional effort that was expected to boost the beneficial effects of testing. The fourth empirical study (Chapter V) was designed to investigate the potential of adapting question difficulty of multiple-choice questions according to learners' knowledge of the learning content and thus create difficulties that are desirable and effective means of retention.

This dissertation is structured as follows: Chapter I will provide the theoretical background of the direct testing effect. Furthermore, relevant research from both laboratory and applied educational contexts will be discussed. On this basis, the rationale for the presented research that investigates direct testing effects in real-world educational contexts will be elaborated. Chapters II to V report the four empirical studies conducted in pursuit of the aims of this dissertation. Chapter VI summarizes the findings of the empirical studies and discusses theoretical and practical implications of this dissertation.

References

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). MIT Press.

Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: The benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(6), 1563–1569. https://doi.org/10.1037/a0017021

Dunlosky, J., & Rawson, K. A. (2015). Practice tests, spaced practice, and successive relearning: Tips for classroom use and for guiding students' learning. *Scholarship of Teaching and Learning in Psychology*, *1*(1), 72–78. https://doi.org/10.1037/stl0000024

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

Dunn, D. S., Saville, B. K., Baker, S. C., & Marek, P. (2013). Evidence-based teaching: Tools and techniques that promote learning in the psychology classroom. *Australian Journal of Psychology*, *65*(1), 5–13. https://doi.org/10.1111/ajpy.12004

Finley, J. R., Benjamin, A. S., Hays, M. J., Bjork, R. A., & Kornell, N. (2011). Benefits of accumulating versus diminishing cues in recall. *Journal of Memory and Language*, *64*(4), 289–298. https://doi.org/10.1016/j.jml.2011.01.006

Glover, J. A. (1989). The "testing" phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology*, *81*(3), 392–399.

Hartwig, M. K., & Dunlosky, J. (2012). Study strategies of college students: Are self-testing and scheduling related to achievement? *Psychonomic Bulletin & Review*, *19*(1), 126–134. https://doi.org/10/b6w3xm

Karpicke, J. D., Butler, A. C., & Roediger, H. L. (2009). Metacognitive Strategies in Student Learning: Do Students Practise Retrieval When They Study on Their Own? *Memory*, *17*(4), 471–479. https://doi.org/10.1080/09658210802647009

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

Kornell, N., & Bjork, R. A. (2007). The promise and perils of self-regulated study. *Psychonomic Bulletin & Review*, *14*(2), 219–224. https://doi.org/10/dbfstw

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

Kornell, N., & Vaughn, K. E. (2016). How retrieval attempts affect learning. In *Psychology of Learning and Motivation* (Vol. 65, pp. 183–215). Elsevier. https://doi.org/10.1016/bs.plm.2016.03.003

Kromann, C. B., Jensen, M. L., & Ringsted, C. (2009). The effect of testing on skills learning. *Medical Education*, *43*(1), 21–27. https://doi.org/10.1111/j.1365-2923.2008.03245.x

McDaniel, M. A., & Little, J. L. (2019). Multiple-choice and short-answer quizzing on equal footing in the classroom: Potential indirect effects of testing. In J. Dunlosky & K. A. Rawson (Eds.), *The Cambridge Handbook of Cognition and Education* (1st ed., pp. 480–499). Cambridge University Press. https://doi.org/10.1017/9781108235631.020

Morehead, K., Rhodes, M. G., & DeLozier, S. (2016). Instructor and student knowledge of study strategies. *Memory*, *24*(2), 257–271. https://doi.org/10/gf2zsb

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*, *4:5*. https://doi.org/10/gf2rp4

Phelps, R. P. (2012). The effect of testing on student achievement, 1910–2010. *International Journal of Testing*, *12*(1), 21–43. https://doi.org/10.1080/15305058.2011.602920

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

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. https://doi.org/10.1111/j.1745-6916.2006.00012.x

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin*, *140*(6), 1432–1463. https://doi.org/10.1037/a0037559

Rummer, R., Schweppe, J., Gerst, K., & Wagner, S. (2017). Is testing a more effective learning strategy than note-taking? *Journal of Experimental Psychology: Applied*, *23*(3), 293–300. https://doi.org/10.1037/xap0000134

Stenlund, T., Jönsson, F. U., & Jonsson, B. (2017). Group discussions and test-enhanced learning: Individual learning outcomes and personality characteristics. *Educational Psychology*, *37*(2), 145–156. https://doi.org/10.1080/01443410.2016.1143087

Struyven, K., Dochy, F., & Janssens, S. (2005). Students' perceptions about evaluation and assessment in higher education: A review. *Assessment & Evaluation in Higher Education*, *30*(4), 325–341. https://doi.org/10.1080/02602930500099102

Toppino, T. C., LaVan, M. H., & Iaconelli, R. T. (2018). Metacognitive control in self-regulated learning: Conditions affecting the choice of restudying versus retrieval practice. *Memory & Cognition*, *46*(7), 1164–1177. https://doi.org/10/gdqczt

Vaughn, K. E., & Kornell, N. (2019). How to activate students' natural desire to test themselves. *Cognitive Research: Principles and Implications*, *4*(1), 35. https://doi.org/10.1186/s41235-019-0187-y

Zeidner, M. (1987). Essay versus multiple-choice type classroom exams: The student's perspective. *The Journal of Educational Research*, *80*(6), 352–358. https://doi.org/10.1080/00220671.1987.10885782

Chapter I

Testing Effects in Real-World Educational Settings: Evidence and Open Questions

The Testing Effect

In their seminal study, Roediger and Karpicke (2006a) investigated the research question whether the testing effect—sporadically studied since 1907 (Witasek, 1907)—would be beneficial in a controlled laboratory setting while using real-world educational material. Participants initially read prose material and subsequently took a practice test on the before studied content and later on a criterial test that measured retention of the studied learning material. To limit unwanted influences the researchers withheld any kind of feedback (i.e., information about the correctness of their given answers or what the correct answers would have been). Thus potential effects of taking a practice test can be interpreted as the results of taking a test alone. Additionally, taking a practice test was compared to restudying the prose material. This design enabled the authors to compare the beneficial effects of practice tests to another learning activity that also implies re-exposure to the learning material. The results were in favor of the testing effect: Although participants in the testing condition only read the prose material once, they outperformed participants that restudied the prose material and thus were exposed to the learning content twice. Consequently, the authors interpreted the findings as practice tests being beneficial for retention as a direct effect of testing because they ruled out possible other explanations, such as mediating processes triggered by taking a practice test. Such mediating processes are referred to as indirect testing effects. Possible indirect testing effects might be that practicing test questions leads to more continuous and overall increased study behavior in expectancy of tests, that feedback as additional exposure to the learning material might present an additional learning opportunity, or that practicing test questions alters meta-cognitive processes (for reviews of indirect testing effects, see McDaniel & Little, 2019; Roediger et al., 2011; Roediger & Karpicke, 2006b). Additionally, anxiety-reducing effects caused by testing might lead to an improved performance in the

criterial test. Furthermore, testing might provide meta-memorial advantages to learners such as identifying learning content that has not been mastered before, which is assumed to improve both the ability to transfer and the organization of knowledge.

With their findings that taking a practice test is beneficial even without feedback and that testing is even superior to other learning activities such as reexposure to the learning information, the works by Roediger and Karpicke (2006b, 2006a) played the most important role to rekindle the interest in the testing effect and spawned an ever growing body of research and (re-)investigation of the testing effect in real-world educational contexts (see Karpicke, 2017, for a review of research between 2006 and 2016).

This dissertation aims to advance the understanding of direct testing effects in educationally relevant learning environments. To this end, I will first highlight theoretical and practical considerations that motivated this research. In the following section, I will summarize findings from investigations of the testing effect in real world educational contexts and review the methodology used in these studies to determine the interpretability of these findings to the current research question. In the last two sections of this chapter, I will present theoretical accounts on the (direct) testing effect and on important moderators, before formulating specific research questions addressed in the studies presented in Chapters II–V.

Direct Testing Effects in Educational Settings

The testing effect is defined as the beneficial effect of taking a practice test as compared to other study activities that are equally time-consuming and reexpose learners to the learning content. Furthermore it has been claimed that the testing effect is also beneficial for retention when no feedback is provided. Beneficial effects of practicing test questions even without feedback have been attributed to the direct effects of testing, which are thought to increase retention of learned information because of processes elicited solely by practicing

the retrieval of this information. This differentiation between a testing effect with subsequent feedback and direct testing effects as result of taking a practice test irrespective of feedback is crucial for both theoretical and practical reasons.

When comparing potential effects, practice tests possibly have on retention of practiced information, I have already stated that providing learners with feedback subsequent to practice might trigger processes that are not triggered when feedback is withheld. In fact it is assumed, that most indirect testing effects are a result of practicing questions and providing test takers with the correct answers (i.e., corrective feedback) whereas informing test takers whether the given answer was correct does not trigger indirect effects of testing (McDaniel & Little, 2019; Pashler et al., 2005).

Additionally to indirect benefits of testing, practice testing with feedback differs from practice testing without feedback in respect of the exposure of learners to the learning content and thus explains the beneficial effects on retention simply by the longer engagement with the learning material. Common paradigms investigating the testing effect first present the to-be learned information in the form of word lists, prose material, or lectures, to learners before they take practice tests or restudy the learning content. Up to this point, exposure to the learning content is equal. Providing feedback in the form of presenting the correct answers to learners means additional exposure to the learning content, often by means of re-presenting learners with restudy material subsequent to practice tests (Rowland, 2014). Current theoretical accounts on the testing effect incorporate the assumption that recalling the correct answer from memory is equivalent to being told the correct answer (Kornell & Vaughn, 2016). Consequently, in terms of exposure, taking a practice test and correctly recalling information is equivalent to receiving feedback on that information. Therefore, learners are presented with the learning material twice when restudying (initial learning +

restudy), whereas when taking a practice test, learners encounter the learning content three times (initial learning + practice test + feedback). This imbalance of exposure to learning content might explain why, when compared directly, practice tests with additional feedback outperforms restudying in studies conducted in both the laboratory (for a meta-analysis, see Rowland, 2014) and in applied educational settings (for a qualitative review, see Moreira et al., 2019). Support for the theoretical claim that an imbalance of exposure to the learning content might be the cause for the superiority of testing with feedback over restudying is supported by laboratory research demonstrating no testing effects when feedback is withheld, but emerging testing effects when feedback is provided (Kang et al., 2007). A similar pattern of results is also observable in applied research (Lipko-Speed et al., 2014), however, reexposure as well as the aforementioned induction of indirect testing effects might explain these findings.

To summarize theoretical reservations about research conducted in both the laboratory and applied educational settings, findings from studies indicating that practice tests with subsequent feedback is superior to restudying might only indicate that restudying learners have not been reexposed to the learning content often enough and thus multiple restudy occasions might be equally effective as taking practice tests. It is therefore necessary to investigate the effects of taking a practice test apart from feedback and reexposure that might come with it.

Another practical consideration that motivated research of testing effects without the provision of feedback concerns the costs of feedback. Feedback consumes educators' and learners' time that could have been spent on additional learning. It has been found that feedback can have negative net effects on learning when total learning time was limited (Hays et al., 2010). Furthermore, it takes time to develop good practice questions and thus

practitioners might not always want to reveal the correct answers because they wish to reuse the same questions in graded examinations.

It has often been argued against the practical importance of the investigation of testing effects without feedback because situations in which learners are tested and do not receive feedback are "[…] exceedingly rare in education: When people cannot think of the answer to the question, they try to figure it out; when they cannot think of the word on the back of a flashcard, they turn the card over; and when a teacher asks a question and the students get it wrong, she tells them the correct answer." (Kornell & Vaughn, 2016, p. 186). It seems consensual among researchers to assume that when learners encounter a question they cannot answer or answer incorrectly, learners experience a need to know the correct answer and educators will want to provide correction. As outlined above, there are scenarios in education where instructors shy away from the costs that are linked to the provision of feedback which ultimately leads to learners—especially in higher education—being responsible for receiving corrective feedback. However, learners are not always willing or able to obtain feedback on their own (Dunlosky & Ariel, 2011). To foreshadow findings from Study 3 presented in this dissertation, even when answering practice questions on learning content that is part of their curriculum, learners did not look up the correct information more often than learners who did not practice questions and when they tried to look it up, they did not find the information that corrected their erroneous response.

To conclude, motivation for the current research question comes from two sources.

For one, from a theoretical perspective it is essential to know whether the testing effect found in applied educational contex is the result of testing or of feedback. To this end, research that investigates direct testing effects in educational contexts is necessary because

direct testing effects—per definition—reflect the effects of testing without the effect of feedback.

For another, investigating the extent and mechanisms behind the direct testing effect in educationally relevant contexts imparts knowledge to educators that can be used in the creation of evidence-based didactic tools and provides students with guidelines how to maximize retention. It should appear intriguing to investigate whether direct benefits are also present in real-world educational settings and thus practically relevant for designing learning environments, especially since direct testing effects can be applied to many educational settings where feedback subsequent to practice tests is not possible because of the outlined reasons.

Reviews of Testing Effects in Real-World Educational Settings

The effects of practice tests are still heavily researched, even after a century of investigations in both the laboratory and the classroom. A query in the Web of Science with "testing effect" or "retrieval practice" as a key term resulted in 142 papers published in the year 2019 thus continuing the sharp rise between the years 2005 and 2015 (Karpicke, 2017). Recent meta-analyses indicate that practicing test questions consistently improves retention with medium effect sizes in laboratory research (Hedges' g = 0.50, Rowland, 2014) and research conducted in university classes (Cohen's d = 0.56, Schwieren et al., 2017). Similar and larger effect sizes have been observed in meta-analyses covering both applied and laboratory research (Cohen's d/Hedges' g between 0.51 and 0.88, Adesope et al., 2017; Phelps, 2012).

With regard to direct testing effects, evidence from meta-analyses seems to indicate that even without feedback, practice tests had beneficial effects on retention ranging between g = 0.39 in laboratory research (Rowland, 2014) and g = 0.60 across laboratory and applied

research (Adesope et al., 2017). Meta-analytic evidence from investigations in psychology classes seems to indicate similar effects of practice tests without feedback (d = 0.47, Schwieren et al., 2017). It thus seems reasonable to assume that practicing test questions in real-world educational settings can have direct effects on retention. However, apart from Rowland's meta-analysis, all meta-analyses included studies that investigated beneficial effects of practice testing irrespective of the control condition and provision of feedback. To investigate whether practice testing is also beneficial as compared to restudying in applied educational settings, some meta-analyses included the control condition as a factor and did find medium to large testing effects (g = 0.51, Adesope et al.,; d = 0.73, Schwieren et al.,). Meta-analyses including studies of practice test in applied educational settings also investigated the effects of feedback as a factor and did find that practice tests are beneficial even without feedback (g = 0.60, Adesope et al.,; d = 0.47, Schwieren et al.,). Although, meta-analyses of research in applied educational settings indicates that practicing testing is superior to restudying and irrespective of feedback, it should be noted that these factors were never combined in these analyses. In studies that were included in these meta-analyses it is very common to provide feedback and in the small body of studies that withheld feedback, often no comparison of practice tests to restudy was employed (see next section for a methodological review). As a result, the finding that practice tests are superior to restudy might only include studies in which participants received feedback subsequent to being tested and the finding that practice tests are beneficial even without feedback might be based solely on studies that did not compare practice testing to an adequate control condition. It thus seems premature to conclude that direct testing effects are practically relevant in real-world educational settings.

Chapter I

A recently published review of research articles on the testing effect in applied educational settings investigated, whether the benefits of practice tests are relevant in applied educational settings (Moreira et al., 2019). In a methodological review, the authors investigated whether studies used learning materials stemming from the actual course or class and thus was eventually examined ("Materials for actual exams?"), what taking a practice test was compared to ("Control condition"), and whether feedback was provided subsequent to testing ("Feedback"). The methodology of all reviewed articles is summarized in Table 1 along with a remark indicating whether taking a practice test produced beneficial effects ("Was retrieval practice overall beneficial?").

Table I.1

Summary of Methodology and Occurrences of Testing Effects in the Review by Moreira et al. (2019)

References	Materials for actual exams?	Control condition	Feedback	Was retrieval practice overall beneficial?
Balch, 1998	Yes	Restudy	Yes	Yes
Batsell et al., 2017	Yes	No activity	No	Yes
Burdo & O'Dwyer, 2015	Yes	Concept mapping vs. No activity	Yes	No
Carpenter et al., 2016	Yes	Copy	Yes	Only for high performance students
Cranney et al., 2009	No	Restudy vs. No activity	Yes	Yes
Daniel & Broida, 2004	Yes	No activity	Yes	Yes
Dirkx et al., 2014	No	Restudy	Yes	Yes
Dobson & Linderholm, 2015	Yes	Restudy or Note taking	Yes	Yes
Goossens et al., 2016	No	Copying	Yes[a]	No
Jaeger et al., 2015	No	Restudy	No	Yes
Kibble, 2007	Yes	No Activity	Yes	Yes

Larsen et al., 2009	No	Restudy	Yes	Yes
Leeming, 2002	Yes	Restudy	Yes	Yes
Lipko-Speed et al., 2014	No	Restudy vs. No activity	Yes vs. No	Yes when feedback provied
Lyle & Crawford, 2011	Yes	Restudy	Yes	Yes
McDaniel et al., 2011	Yes	No activity	Yes	Yes
McDaniel et al., 2013	Yes	No activity	Yes	Yes
McDermott et al., 2014	Yes	Restudy	Yes	Yes
McDaniel et al., 2011	Yes	Restudy vs. No activity	Yes	Yes
Ramraje & Sable, 2011	No	No activity	No	Yes
Vojdanoska et al., 2010	No	No activity	Yes vs. No	Yes
Wiklund-Hörnqvist et al., 2014	No	Restudy	Yes	Yes

Note. [a]In the review of the study by Goosens et al. 2016, Moreira et al. 2019 state that no feedback was administered to students. However, in the light of the following passage, I judged the study to use feedback: "After having completed the first round, the children checked their performance with an answer sheet and corrected wrong answers." (Goossens et al., 2016, p. 6)

Out of 23 reviewed articles, only two studies investigated the combination of withholding feedback subsequent to testing and comparing practice testing to an appropriate control condition. Of these two, only one study did find practice tests to outperform restudying (Jaeger et al., 2015), whereas the other did find a superiority of practice test when feedback was given subsequent to testing but not when feedback was withheld (Lipko-Speed et al., 2014). However, applicability of these two studies to real-world educational contexts is limited because they investigated the testing effect with actual pupils in classrooms but did not use learning content from the curriculum. Instead, pupils learned from texts covering general knowledge or learned definitions of scientific key concepts.

It is a merit of the study by Jaeger and colleagues (2015) to demonstrate the benefits of testing without feedback—and thus most likely, direct testing effects—in a sample of third grade pupils, however usage of learning content that is not part of the curriculum neglects the problem that learning in school differs in terms of motivation, personal involvement, and effort from learning only for the purpose of participating in a psychological or educational study. These differences pose a threat to the external validity of this study and limits its generalizability to the testing effect in actual educational settings.

To summarize, whereas findings from laboratory research indicate the existence of direct testing effects, it is unclear whether these effects are observable and practically relevant in real-world educational contexts. The only literature review investigating methodological approaches in the research of the testing effect came to the conclusion that methodology differs in the investigation of the testing effect in real-world educational contexts and that the investigation of practice testing without feedback is rarely compared to an appropriate control condition. The sole study finding practicing test being superior to

restudying suffered from limited generalizability due to its usage of non-curricular learning content.

Moreira and colleagues (2019) showed that, although these findings point toward the same direction, there are substantial differences in methodology. In an attempt to find more evidence for the presence of direct testing effects in real-world educational contexts, I reviewed the relevant literature with respect to the methodology and findings similar to the one authored by Moreira and colleagues (2019). I did an additional review because of two reasons. First, by the time I wrote this dissertation, Moreira and colleagues' literature search dated back approximately 24 months. As stated initially, testing effects in applied educational settings are heavily researched and the amount of publications seems to increase each year with 142 papers published in the year 2019, thus there might be a large number of recent publications on the subject.

Second, it is apparent that Moreira and colleagues (2019) did not include some studies from the meta-analyses on the subject (Adesope et al., 2017; Schwieren et al., 2017). One possible explanation might be that in some studies that were subject to meta-analyses the term "clicker" is used for practice tests in educational settings. Furthermore the restriction to use only the terms "applied" and "classroom" might seem too narrow to find all relevant researches.

I therefore conducted a literature search with the key terms "'testing effect' OR 'retrieval practice' OR 'test-enhanced learning' AND ('class*' OR 'university' OR 'appl*' OR 'educat*' OR 'clicker')" resulting in 765 publications. This literature search was combined with manual search of all publications that were subject to the meta-analyses that covered applied research on the testing effect (i.e., Adesope et al., 2017; Schwieren et al.,

2017). Relevant publications were selected according to the inclusion criteria by Moreira and colleagues:

"(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 educational institutions." (Moreira et al., 2019, p. 3). Exclusion resulted in 24 publications not already covered by Moreira and colleagues meeting the criteria (Table 2).

Table I.2

Summary of Methodology and Occurences of Testing Effects in the Present Review

References	Materials for actual exams?	Control condition	Feedback	Was retrieval practice overall beneficial?
Bell et al., 2015	Yes	No activity	Yes	No
Bing, 1984	No	No activity vs. Structured restudy vs. Unstructured restudy	No	No
E. L. Bjork et al., 2014	Yes	No activity	No	Yes
Carpenter et al., 2009	Yes	Restudy vs. No activity	Yes	Yes
Downs, 2015	Yes	No activity	Yes vs. No	No, irrespective of feedback
Foss & Pirozzolo, 2017	Yes	No activity	Yes	No
Francis et al., 2020	Yes	No activity	Yes	For retrieval-based concept mapping but not for practicing multiple-choice tests
Johnson & Kiviniemi, 2009	Yes	No activity	No	Yes
Kelley et al., 2019	Yes	No activity	Yes	Yes
Khanna, 2015	Yes	No activity	Yes	Yes

Kromann et al., 2009	Yes	Practicing skill	Yes	Yes
Leggett et al., 2019	No	Restudy	Yes	Yes
Marsh et al., 2012	No	No activity	Yes vs. No	Yes, irrespective of feedback
Mayer et al., 2009	Yes	No activity	Yes	Only for questions answered via clicker response system
McDaniel et al., 2007	Yes	Restudy vs. No activity	Yes	Yes
McDaniel et al., 2012	Yes	Restudy	Yes	Yes
Palmer et al., 2019	Yes	Rewatching class recording	No	No
Rummer et al., 2019	Yes	Open-book retrieval practice	No	Yes
Shapiro & Gordon, 2012	Yes	No activity	Yes	Yes
Stenlund et al., 2017	No	Group discussion	Yes	Yes
Stenlund et al., 2016	No	Restudy	Yes	Yes
Thomas et al., 2020	Yes	Restudy	Yes vs. No	Yes, irrespective of feedback
Trumbo et al., 2016	Yes	Restudy	Yes	Yes

Welch, 2019	Yes	No activity	Yes	Only when repeating pratice test twice as compared to one test

Considering the methodology, the observations made by Moreira and colleagues also apply to relevant publications that came up in my literature search: Methodology differs greatly among studies investigating the testing effect in real-world educational settings, however studies investigating testing effects without feedback in comparison to an appropriate control condition are scarce and findings are inconsistent. Four studies investigated the effects of taking practice tests without feedback as compared to an appropriate control condition. The first study, authored by Bing (1984) investigated whether retention of prose material could be increased by different practice tests subsequent to reading. Taking practice tests had no effect on retention when compared to the restudy conditions. Similar to the presented studies by Lipko-Speed and colleagues (2014) and Jaeger and colleagues (2015), used learning material was not part of the curriculum and thus generalizability of these findings to real-world educational contexts is limited. The second study, authored by Palmer and colleagues (2019) however, did investigate the effects of taking practice tests without feedback subsequent to attending a lecture as compared to rewatching the lecture. This study did also find no difference between taking a practice test and rewatching a recorded lecture. The third study, authored by Thomas and colleagues (2020) also investigated the effects of practice tests in real-world university classes. To do so, the researchers redesigned an existing psychology class to investigate the effects of 10-minute practice tests or restudying subsequent to every 20 minutes of lecture. In three unit exams taking practice tests with and, most notably as well as without feedback promoted retention more than did restudying. This study is—to my knowledge—among the first to find

beneficial effects of multiple-choice practice tests without feedback in comparison to restudying in a real-world educational context (see also Chapter IV). However, it is unclear whether this benefit of testing is a result of a direct testing effect. For one, students' performance in the practice tests counted towards their course grade. For another, manipulation of the 10-minute practice session subsequent to the lecture occurred within students and comprised of either taking practice tests, taking practice tests with feedback in the next class meeting, or restudying. Every 4 weeks, students experienced another practice strategy, whereas each week included two class meetings and each class meeting comprised two lectures and the corresponding test or restudy sessions. In other words, students could easily infer upcoming tasks following each lecture from the practice session they experienced. This ease in predicting whether a lecture is followed by a graded practice test or a restudying opportunity might pose a threat to the assumption that the found benefit reflects direct testing effects because it might change learning behavior: Expectancy of a graded test might maximize effort and motivation in lectures as compared to restudying. The fourth study by Rummer and colleagues (2019) investigated practicing short-answer questions without feedback in the last minutes of a university course to an almost identical condition in another course: Participants answered short-answer questions while allowing them to use whatever learning materials they thought helpful to answer the questions (open-book condition). The authors found a significant increase in retention whenever participants were not allowed to use additional material (closed-book condition) as compared to the open-book condition. Comparison of these conditions in an existing learning environment is valuable to in the investigation of direct testing effects because the sole difference between these two conditions is that to answer the practice questions in the closed-book questions participants have to retrieve information from memory whereas in the open-book condition they do not

have to. However, there are also some limitations in the interpretability of the results. For one, it is stated that the third author was responsible for teaching the two courses whereas manipulation of retrieval practice varied between courses and that the third author was not entirely blind to the manipulation and might have—knowingly or unknowingly—affected the outcomes of the study. Additionally, interpretability of this study might suffer from an issue related to the one present in the study by Thomas et al. (2020): Participants might have anticipated the task that was required in the last minutes of the course and adjusted their behavior accordingly by either preparing themselves to answer practicing questions with or without consulting their learning material.

Thus the research question, whether direct testing effects are relevant in real-world educational contexts by means of increasing retention in comparison to other strategies that foster retention is currently underexplored in the scientific literature and yields inconsistent findings.

Inconsistent findings have also been observed in simulated classroom studies, i.e. laboratory studies investigating testing effects without feedback and in comparison to an appropriate control condition but with real-world educational materials. Some studies find beneficial effects of taking practice tests (Einstein et al., 2012; Nungester & Duchastel, 1982), some find differential effects depending on the question format (Butler & Roediger, 2007), while others find no beneficial effects at all (Kang et al., 2007).

These found inconsistencies can partly be explained by theoretic accounts on the testing effect that incorporate moderators always present when taking practice tests. These accounts are the subject of the next section.

Theoretical Accounts on Memory and the Testing Effect

In the current literature on the testing effect, several theoretic accounts on this phenomenon are discussed. It is noteworthy that these accounts do not necessarily contradict each other or are mutually exclusive. Furthermore there are theoretic accounts on the mechanisms of the benefits of practice tests over restudying, that explain why this phenomenon occurs but make little to no assumptions or predictions of potential moderators in applied educational settings when no corrective feedback is provided. These accounts include the elaborative retrieval hypothesis (Carpenter, 2009, 2011), the search set theory (Grimaldi & Karpicke, 2012), and the episodic context account (Lehman et al., 2014).

In the following I will first present a theory of memory that is applicable to a wide array of memory phenomena including the testing effect and then I will turn to more specific theoretical accounts on the testing effect that rely on the theory presented first. I will focus on theoretical accounts that make predictions about the presence and the extent of the testing effect without the provision of corrective feedback and in comparison to control conditions. Presented theories should be able to both explain obtained findings in the research literature as well as predicting ways to improve the effectiveness of practice tests in applied educational settings.

The New Theory of Disuse

The new theory of disuse (R. A. Bjork & Bjork, 1992) is a versatile theory of memory that aims at explaining why information that has been presented on a single occasion is sometimes forgotten after hours and sometimes lasts in memory for a lifetime. The prospect of answering this question is very relevant for many aspects that involve human memory most notably in the fields of learning and education and consequently practical implications have been derived from this theory (R. A. Bjork & Bjork, 2006). This theory is

furthermore important as it often serves as a referential frame for contemporary theoretical accounts on the testing effect and by itself can be seen as the theoretical basis of the desirable difficulties framework (R. A. Bjork, 1999). The main claim of the new theory of disuse is that recall from memory is a crucial modifier for the retention of memory content (R. A. Bjork & Bjork, 1992): The novel concept that presents the core of this theory is that memory content is characterized by two "strengths": Storage strength—a latent variable—reflects how well memory content is learned in general, while retrieval strength reflects the current ease with which memory content can be retrieved from memory and mainly determines the outcomes of memory tests and exams. Whenever information is studied storage strength and retrieval strength accumulate, however it is assumed that active processing of studied information that require a successful retrieval from memory—such as practicing test questions and other desirable difficulties—is associated with more gains in storage strength and retrieval strength than (re)studying is. It is noteworthy to add that future gains in storage strength are higher the lower retrieval strength and storage strength are currently, which represents the observation that the increments in learning are biggest when relatively little is known about a subject (low storage strength) and learning content is not easy to be accessed or activated in memory (low retrieval strength). Similar to gains in storage strength, future retrieval strength increases more with lower current retrieval strength, but also with higher storage strength. The authors state that the latter draws on the observation that well-learned content can be relearned rapidly, enabling quick access to once learned information, with significant effort savings compared to its initial learning. Similarly, it is assumed that whenever retrieval strength decreases (see next paragraph) future decrease will be the less the higher storage strength is currently.

Concerning the forgetting of learning content, it is assumed that once accumulated storage strength is never lost, however memory content can become inaccessible because of low retrieval strength. In contrast to storage strength, retrieval strength is assumed to be limited by both the amount of learning content and the discriminatory attributes of the learning content: With every information that is studied, previous information would suffer from decreased retrieval strength and it will become harder to discriminate the correct information in memory from other information resulting in an additional decrease in retrieval strength.

Assumptions made within this theory are in line with the idea of a testing effect for items that have been studied once and when no feedback is administered subsequent to testing: Whenever once learned material is subject to practice tests, both storage and retrieval strength receive a gain that surpasses the gain received from restudying this material. Additionally, the theory assumes higher increments in both retrieval and storage strength following practice tests to be the result of lower retrieval strength when taking the practice test. In other words, practice test questions that are harder to answer are assumed to profit learning more.

The Bifurcation Model

The main goal of the bifurcation model (Halamish & Bjork, 2011; Kornell et al., 2011) is to provide an explanation why practice tests without feedback are superior to restudying in general (Roediger & Karpicke, 2006b) and to also include findings which suggest that restudying is more effective than practice tests are (e.g., Wheeler et al., 2003).

The explanations brought forward by the bifurcation model draw on the ideas of the new theory of disuse (R. A. Bjork & Bjork, 1992), mainly on the assumption that correctly recalling information from memory increases the retrieval strength of that piece of

information more than restudying does. To understand why restudying has sometimes been found to be superior or equal to practicing test questions, the model highlights the implications of information that has been practiced unsuccessfully: Whereas successfully recalled information receives a major boost in retrieval strength and restudied information receives a small boost in retrieval strength, the bifurcation model assumes in accordance with the new theory of disuse that unsuccessful retrieval not only results in no boost but also in additional decrease in retrieval strength due to the assumption that non-practiced information is subject to slow decrease in retrieval strength, commonly described as forgetting.

Another assumption of the bifurcation model is that once studied information is normally distributed concerning their associated retrieval strength, which means that most information has medium retrieval strength with half of the information having higher and the other half having lower retrieval strength, respectively (Kornell et al., 2011).

The bifurcation model‘s main focus is on the distribution of a given set of information.

For practice test questions as well as for criterial tests questions, the bifurcation model introduces a recall threshold which must be surpassed by the retrieval strength of the associated piece of information in order to answer a question correctly. This threshold is a key assumption of this model because this threshold can be seen as the criterion that renders practicing test questions either successful or unsuccessful. It is assumed that whenever a practice question is encountered one of two outcomes are possible: Whenever the retrieval strength associated to the information subject of a practice test question is above the threshold, the necessary piece of information can be retrieved from memory and this piece of information receives a major boost to its retrieval strength. However, when retrieval strength associated to the information subject of a practice test question is below the threshold, the

piece of information cannot be retrieved and thus receives no boost in retrieval strength. Answering practice test questions results in a bifurcated distribution of retrieval strength because all information above a certain criterion of retrieval strength receive additional boosts in retrieval strength whereas all information that have not been retrieved because the retrieval strength was too low in the first place, will receive no gains in retrieval strength. This consequently leads to a boost in retrieval strength for information that already displayed highest retrieval strength resulting in the creation of a gap between these two groups of information.

Contrary to the bifurcated distribution as a result of retrieval practice, it is assumed that restudied information will always benefit from a gain in retrieval strength. This gain is independent of the initial retrieval strength of the information and thus increases retrieval strength of the entire distribution associated with the information set.

On the basis of these assumptions, the bifurcation model can explain positive, negative and no effects of practicing test questions even without corrective feedback by the amount of a given set of information (e.g., bits of information that were part of a lecture students just visited) that surpasses the recall threshold set by the practice test questions: Whenever the threshold can be surpassed by the majority of information, this majority is most likely remembered in a later criterial test because it received boosts to its retrieval strength that are essential to surpass the threshold set by the criterial test questions. Restudied information received only a minor boost in retrieval strength which might result in some information surpassing the threshold set by criterial test questions, however the amount of surpassed thresholds and thus correct answers in the criterial test will be higher for correctly answered test questions. On the basis of the criterial test, practicing test questions produced

more correct answers in the criterial test than restudying did and thus practicing test question can be seen as the more effective learning strategy in this scenario.

However, when practicing test questions and information‘s retrieval strength can rarely surpass the threshold, practicing retrieval questions has no benefit in regard to retrieval strength and can thus be outperformed by restudying, where all items receive at least a small gain in retrieval strength, resulting in the absence of a testing effect or even in negative testing effects.

The bifurcation model furthermore aims at explaining why under some circumstances testing effects are more likely to occur than under others, by entering the time interval between the (last) practice of the information and the criterial test (i.e., retention interval) into the model (Halamish & Bjork, 2011; Kornell et al., 2011). It is a common finding that restudying benefits retention more than practicing test questions does after short retention intervals, whereas the pattern of results and thus the typical testing effect only emerges after longer retention intervals (for a review, see Roediger & Karpicke, 2006b). Given that forgetting and thus a slow decrease in retrieval strength is independent of acquired retrieval strength, it also applies to both restudied information as well as to information that has been practiced by answering test questions. However, a decrease of retrieval strength of restudied information has more profound consequences for retrieval strength in a later criterial test, because the gain in retrieval strength that has been elicited by restudying was small and thus the longer the retention interval grows, the smaller the net value of restudying will be.

According to the bifurcation model, the same effects of retention intervals apply when information was subject to practice questions, however, answering test questions correctly elicited a major boost in retrieval strength and thus decrease in retrieval strength

starts from a higher value. Reconsidering that in order to be deemed more effective in comparison to restudying, practicing test questions is required to produce more correctly answered criterial test questions—and thus more information with retrieval strengths above the criterial test threshold—than restudying. Due to the minor boost in retrieval strength for all restudied information, a large proportion of information can surpass the threshold set by an immediate criterial test. With longer retention intervals, more and more information falls below the threshold of the criterial test. The beneficial effects of answering test questions correctly might not apply to as much information as the benefits of restudying did, but the effects are bigger in magnitude and thus it will take longer retention intervals for tested information to fall under the threshold of the criterial test. As a result the bifurcation model predicts that in the short run, restudying might seem as the more effective strategy but the testing effect will prevail in the long run.

The Retrieval Effort Hypothesis

The second theoretical account that makes predictions about the testing effect in educational settings presented here also draws on the new theory of disuse but mainly focuses on the beneficial effects of successful retrieval of information with varying amounts of retrieval strength. The new theory of disuse assumes that the lower the retrieval strength of an information that has been successfully recalled is, the higher the gain in retrieval strength fort that information is. This assumption is similar to the main claim of the desirable difficulties framework (R. A. Bjork, 1994) that—on the basis of findings of almost a century of research—proposes the idea that difficult but solvable practice fosters long-term retention more than easy practice does. The retrieval effort hypothesis (Pyc & Rawson, 2009) proposes to explain and predict the effects of difficulty in practicing test questions and consequently states that difficult but correctly answered practice tests are better for retention than easier

correctly answered practice tests are. It should be noted, that this hypothesis solely focusses on successful attempts to answer practice questions and that it assumes that feedback following successful practice has no additional beneficial effect (for evidence, see Pashler et al., 2005). Although a definition of difficulty is never given in this theoretical account, difficulty follows the idea of the inverse of retrieval strength and other assumptions of the new theory of disuse: Difficulty rises with the time that passes between the first study of information and subsequent practice of that information (Pyc & Rawson, 2009) and with the weakened associations between two concepts (Carpenter, 2009). Most importantly for the application of testing effects in real-world educational contexts, difficulty within a set of information might simply arise as a characteristic of the individual information, as for example, some lecture content will almost always be remembered at the end of a lecture because learners have a strong emotional relation to it or it bears great personal relevance for them (Vaughn et al., 2013).

Practical Implications of Theoretical Accounts on the Testing Effect

To summarize the two theoretical accounts on the testing effect, the bifurcation model assumes that whether a testing effect emerges without the provision of feedback depends on whether test questions can be answered—and thus be retrieved—correctly and how long the retention interval is. The amount of successful retrievals is dependent on the level of the information‘s retrieval strength that surpasses the threshold set by the practice test. Retrieval strength in combination with the recall threshold determines the retrievability of learning content in actual educational contexts. Evidence suggests that retrieval success depends on question difficulty and is highest for easy items (Kornell & Vaughn, 2016). Easy items should thus increase success rates and every correctly recalled item increases retention as compared to restudied items. For information such as learning content presented in a lecture, retrievability can thus be seen as the inverse of psychometric question difficulty, with

higher retrievability resulting in more correct retrievals and thus correct answers in practice tests and hence a lower difficulty and vice versa. Consequently, practice test question difficulty plays a large role in educational settings because the bifurcation model assumes them to be beneficial, only when they are easy enough to promote a certain amount of retrievability.

The retrieval effort hypothesis also assumes that in order to be more effective as compared to restudying, practicing test questions should include successful retrievals from memory and thus answering practice questions correctly. Furthermore, regarding information that has been retrieved successfully, this hypothesis assumes practice tests to be the more effective, the more difficult the retrieval is. In educational contexts however, difficulty of practice questions might be detrimental or beneficial for practice test questions because difficulty of practice questions is associated with (a) whether they are answered correctly and (b) their beneficial effect on retention if answered correctly. Answering easy practice questions might thus result in more successful retrieval attempts than answering hard practice questions but in comparison to restudying the former might result in less benefit on retention than the latter.

Thus the assumed role of question difficulty in testing in real-world educational settings is different for the presented theoretical accounts on the testing effect. Whereas the bifurcation model assumes that the easiest practice test questions will lead to maximized testing effects, the retrieval effort hypothesis assumes that difficulty has to be balanced in order to promote successful and effortful retrieval.

Different theoretical accounts might also explain the inconsistent findings from both actual and simulated classrooms. Easier questions have been associated with higher testing effects in the bifurcation model. This model would provide a possible explanation why a

testing effect was observed for short-answer questions without additional feedback in the study by Butler and Roediger (2007) with practice question performance of 68% but not in the study by Kang and colleagues (2007) with practice question performance of 54%.

As stated in the introduction, the format in which the practice test takes place alters the difficulty of the practice questions. The differing assumptions about the role of question difficulty can also be extended to predictions about different test formats. Rawson and Zamary (2019) derived assumptions on the basis of the bifurcation model and the retrieval effort hypothesis for the difference between recognition and cued-recall as well as free-recall questions. Recognition tasks are assumed to foster more successful responses and consequently successful retrieval from memory. However, recognition has been associated with less effort, less elaborative retrieval and less complete retrieval than cued-recall questions such as short-answer questions (Carpenter & Delosh, 2006; Glover, 1989). The core assumption of the bifurcation model is that more successful retrievals will ultimately result in higher testing effects and thus multiple-choice questions—that are assumed to rely heavily on recognition—should foster more successful retrieval and thus a greater gain than other question formats. However, the retrieval effort hypothesis assumes greater gains to arise from question formats that are more difficult to answer and require more retrieval effort but consequently will also involve less retrieval success when practicing these questions. Consequently, within the latter theoretical account, multiple-.choice questions are assumed to be least effective in eliciting testing effects.

Additional to theoretical accounts that mainly focus on test question difficulty, the benefit of differing practice test question difficulty in real-world educational setting is not solely dependent on characteristics of the practice test but also on learners' abilities and knowledge in other domains. This claim can be backed by research indicating that learners'

individual differences modulate the testing effect dependent on practice test difficulty (Fiechter & Benjamin, 2017; Minear et al., 2018). Differing abilities among learners have shown different effect patterns as a result of practicing test questions that varied in their difficulty: Learners with high abilities have shown testing effects for difficult questions whereas learners with low ability levels only profited from practice tests as compared to restudying when questions were easy (Minear et al., 2018). Further research has shown that lowering the difficulty of practice questions was found to elicit testing effects whereas the practice of questions with their original difficulty lead to similar effects on retention thanrestudying did(Fiechter & Benjamin, 2017). Consequently, to understand the contribution of practice question difficulty totesting effects without feedback in actual educational settings, explanations should include learners' abilities.

Research Questions

In the preceding sections, I outlined the motivation to investigate direct testing effects in real-world educational contexts and described the current state of research on this matter. Furthermore, I presented theoretical accounts that make predictions about the effectiveness of practicing test questions without corrective feedback as part of everyday educational settings. These theoretical accounts agree on the assumption that test question difficulty and the strongly associated retrievability of information from memory are important factors to consider when using practice tests without feedback. It has been argued that differing test question difficulty might explain why test formats (e.g., multiple-choice tests, short-answer tests) sometimes differ in their beneficial effect on retention (Butler & Roediger, 2007) and thus investigations of different practice test formats have implications for employing testing in real-world educational contexts. The aim of this dissertation is to advance a more comprehensive understanding of direct testing effects in applied educational

contexts, its moderators and ways to improve their effectiveness in these contexts by empirically exploring the relationship between practice question difficulty, direct testing effects, and a subset of related research questions.

The main goal of the first study (Chapter II) was to investigate whether direct testing effects can be applied fruitfully to existing educational settings which means that practicing test questions about actual learning content improves retention more than restudying learning content does and that this increase is also present over time periods that reflect time periods in actual educational contexts. Furthermore, this study was aiming to investigate under what conditions direct testing effects occur in existing educational settings. It has been outlined that theoretical accounts on the testing effect assume practice question difficulty to be a key moderator of direct testing effects. Apart from individual differences between practice questions it has been argued that different question formats systematically differ in their difficulty: Answering multiple-choice questions has often been deemed less difficult than answering short-answer questions. Study 1 consequently investigated the role of two practically relevant factors that are derived from these theoretical considerations. In this study, students attended a lecture and then practiced either short-answer or multiple-choice questions about the lecture's content or restudied the lecture's content, thus investigating the moderating role of practice question format on direct testing effects. Additionally the study investigated the moderating role of naturally occurring differences in practice question difficulty in the form of retrievability of individual parts of the lecture's content, operationalized by calculating the difficulty of the practice questions.

Study 2 (Chapter III) was a conceptual replication of Study 1 in its attempt to investigate direct testing effects and practically important moderators. Study 2 however differed in the way multiple-choice questions were operationalized in order to render

practicing this question format more effective as compared to restudying. It has previously been stated that multiple-choice questions are associated with comparably low question difficulty because to answer correctly in this question format, respondents merely need to recognize the single correct response option instead of recalling the correct information from memory. In an attempt to increase practice question difficulty, multiple-choice questions in Study 2 were designed in a way that encouraged respondents to process all response options before answering. Investigating the effects of these modified multiple-choice questions in comparison to restudying should provide more evidence for the applicability of multiple-choice questions as means to elicit direct testing effects in real-world educational settings.

Study 3 (Chapter IV) further investigated under which difficulty conditions practicing multiple-choice questions is most beneficial in actual university classes. Moreover this study featured an additional factor that is supposed to reflect the moderation of the testing effect in the real world, namely learners‘ feedback behavior. In the light of the finding that feedback following tests appears only to be beneficial for retention, if learners answer practice questions incorrectly (Pashler et al., 2005), it seems efficient to provide learners with corrective feedback only when they need it. The approach to feedback in Study 3 is that learners judge themselves the need for corrective feedback subsequent to answering practice questions. Study 3 explored the research question which factors influence feedback behavior when students practice test questions in an actual university course and how feedback behavior in turn influences beneficial effects of taking practice tests. Furthermore, this study aimed at investigating the interaction of question difficulty and feedback behavior on retention following practice tests in real-world educational settings.

Study 4 (Chapter V) also investigated potential ways to render practice of multiple-choice questions more beneficial when employed in the university classroom. The approach

of Study 4 however, entirely focused on practice question difficulty and its potential to enhance direct testing effects by drawing on the idea that with increasing the difficulty of solvable practice questions, the testing effect will increase as well.

To verify the assumption that answering practice questions tailored to learners‘ ability levels in real-world educational contexts is effective/ beneficial, Study 4 presents a new procedure that provides an increasing assistance to learners practicing multiple-choice questions by successively eliminating incorrect response options until the question is answered correctly. It should be noted that in this approach no feedback was given subsequent to answering practice questions apart from informing participants whether their answer was correct and which incorrect response options had already been eliminated.

The main goal of Study 4 is to compare this novel procedure with standard procedures of retrieval practice and then examine the potential of this adaptive testing procedure for complex learning content in an actual university setting.

The research questions of all four empirical studies of this dissertation are further elaborated in the following chapters.

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

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

Bing, S. B. (1984). Effects of testing versus review on rote and conceptual learning from prose. *Instructional Science*, *13*(2), 193–198. https://doi.org/10.1007/BF00052385

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. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). MIT Press.

Bjork, R. A. (1999). Assessing our own competence: Heuristics and illusions. *Attention and Performance*, *17*.

Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In A. F. Healy, S. M. Kosslyn, & R. M. Shiffrin (Eds.), *From learning processes to cognitive processes: Essys in Honor of William K. Estes* (1992-97939-014; pp. 35–67). Lawrence Erlbaum Associates, Inc.

Bjork, R. A., & Bjork, E. L. (2006). Optimizing treatment and instruction: Implication of a new theory of disuse. In L-G. Nilsson & N. Ohta, *Memory and society: Psychological perspectives*. Psychology Press.

Burdo, J., & O'Dwyer, L. (2015). The effectiveness of concept mapping and retrieval practice as learning strategies in an undergraduate physiology course. *Advances in Physiology Education*, *39*(4), 335–340. https://doi.org/10.1152/advan.00041.2015

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. (2009). Cue strength as a moderator of the testing effect: The benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(6), 1563–1569. https://doi.org/10.1037/a0017021

Carpenter, S. K. (2011). Semantic information activated during retrieval contributes to later retention: Support for the mediator effectiveness hypothesis of the testing effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(6), 1547–1552. https://doi.org/10.1037/a0024140

Carpenter, S. K., & Delosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory & Cognition*, *34*(2), 268–276. https://doi.org/10.3758/BF03193405

Carpenter, S. K., Lund, T. J. S., Coffman, C. R., Armstrong, P. I., Lamm, M. H., & Reason, R. D. (2016). A classroom study on the relationship between student achievement and retrieval-enhanced learning. *Educational Psychology Review*, *28*(2), 353–375. https://doi.org/10.1007/s10648-015-9311-9

Carpenter, S. K., Pashler, H., & Cepeda, N. J. (2009). Using tests to enhance 8th grade students' retention of U.S. history facts. *Applied Cognitive Psychology*, *23*(6), 760–771. https://doi.org/10.1002/acp.1507

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

Daniel, D. B., & Broida, J. (2004). Using web-based quizzing to improve exam performance: Lessons learned. *Teaching of Psychology*, *31*(3), 207–208. https://doi.org/10.1207/s15328023top3103_6

Dirkx, K. J. H., Kester, L., & Kirschner, P. A. (2014). The testing effect for learning principles and procedures from texts. *The Journal of Educational Research*, *107*(5), 357–364. https://doi.org/10.1080/00220671.2013.823370

Dobson, J. L., & Linderholm, T. (2015). 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

Downs, S. D. (2015). Testing in the college classroom: Do testing and feedback influence grades throughout an entire semester? *Scholarship of Teaching and Learning in Psychology*, *1*(2), 172–181. https://doi.org/10.1037/stl0000025

Dunlosky, J., & Ariel, R. (2011). Self-Regulated Learning and the Allocation of Study Time. In *Psychology of Learning and Motivation* (Vol. 54, pp. 103–140). Elsevier. https://doi.org/10.1016/B978-0-12-385527-5.00004-8

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

Fiechter, J. L., & Benjamin, A. S. (2017). Diminishing-cues retrieval practice: A memory-enhancing technique that works when regular testing doesn't. *Psychonomic Bulletin & Review*. https://doi.org/10.3758/s13423-017-1366-9

Foss, D. J., & Pirozzolo, J. W. (2017). Four semesters investigating frequency of testing, the testing effect, and transfer of training. *Journal of Educational Psychology*. https://doi.org/10.1037/edu0000197

Francis, A. P., Wieth, M. B., Zabel, K. L., & Carr, T. H. (2020). A classroom study on the role of prior knowledge and retrieval tool in the testing effect. *Psychology Learning & Teaching*, *19*(3), 258–274. https://doi.org/10.1177/1475725720924872

Glover, J. A. (1989). The "testing" phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology*, *81*(3), 392–399.

Goossens, N. A. M. C., Camp, G., Verkoeijen, P. P. J. L., Tabbers, H. K., Bouwmeester, S., & Zwaan, R. A. (2016). Distributed practice and retrieval practice in primary school vocabulary learning: A multi-classroom study. *Applied Cognitive Psychology*, *30*(5), 700–712. https://doi.org/10.1002/acp.3245

Grimaldi, P. J., & Karpicke, J. D. (2012). When and why do retrieval attempts enhance subsequent encoding? *Memory & Cognition*, *40*(4), 505–513. https://doi.org/10.3758/s13421-011-0174-0

Halamish, V., & Bjork, R. A. (2011). When does testing enhance retention? A distribution-based interpretation of retrieval as a memory modifier. *Journal of Experimental Psychology. Learning, Memory & Cognition*, *37*(4), 801–812. https://doi.org/10.1037/a0023219

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

Jaeger, A., Eisenkraemer, R. E., & Stein, L. M. (2015). Test-enhanced learning in third-grade children. *Educational Psychology*, *35*(4), 513–521. https://doi.org/10.1080/01443410.2014.963030

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

Kang, S. H. K., McDermott, K. B., & Roediger, H. L. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology*, *19*(4/5), 528–558. https://doi.org/10.1080/09541440601056620

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

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 Psychology*, *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

Kibble, J. (2007). Use of unsupervised online quizzes as formative assessment in a medical physiology course: Effects of incentives on student participation and performance. *Advances in Physiology Education*, *31*(3), 253–260. https://doi.org/10.1152/advan.00027.2007

Kornell, N., Bjork, R. A., & Garcia, M. A. (2011). Why tests appear to prevent forgetting: A distribution-based bifurcation model. *Journal of Memory and Language*, *65*(2), 85–97. https://doi.org/10.1016/j.jml.2011.04.002

Kornell, N., & Vaughn, K. E. (2016). How retrieval attempts affect learning. In *Psychology of Learning and Motivation* (Vol. 65, pp. 183–215). Elsevier. https://doi.org/10.1016/bs.plm.2016.03.003

Kromann, C. B., Jensen, M. L., & Ringsted, C. (2009). The effect of testing on skills learning. *Medical Education*, *43*(1), 21–27. https://doi.org/10.1111/j.1365-2923.2008.03245.x

Larsen, D. P., Butler, A. C., & Roediger, H. L. (2009). Repeated testing improves long-term retention relative to repeated study: A randomised controlled trial. *Medical Education*, *43*(12), 1174–1181. https://doi.org/10.1111/j.1365-2923.2009.03518.x

Leeming, F. C. (2002). The exam-a-day procedure improves performance in psychology classes. *Teaching of Psychology*, *29*(3), 210–212. https://doi.org/10.1207/S15328023TOP2903_06

Leggett, J. M. I., Burt, J. S., & Carroll, A. (2019). Retrieval practice can improve classroom review despite low practice test performance. *Applied Cognitive Psychology*, acp.3517. https://doi.org/10/gf2rp2

Lehman, M., Smith, M. A., & Karpicke, J. D. (2014). Toward an episodic context account of retrieval-based learning: Dissociating retrieval practice and elaboration. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(6), 1787–1794. https://doi.org/10.1037/xlm0000012

Lipko-Speed, A., Dunlosky, J., & Rawson, K. A. (2014). Does testing with feedback help grade-school children learn key concepts in science? *Journal of Applied Research in Memory and Cognition*, *3*(3), 171–176. https://doi.org/10.1016/j.jarmac.2014.04.002

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

Marsh, E. J., Fazio, L. K., & Goswick, A. E. (2012). Memorial consequences of testing school-aged children. *Memory*, *20*(8), 899–906. https://doi.org/10.1080/09658211.2012.708757

Mayer, R. E., Stull, A., DeLeeuw, K., Almeroth, K., Bimber, B., Chun, D., Bulger, M., Campbell, J., Knight, A., & Zhang, H. (2009). Clickers in college classrooms: Fostering learning with questioning methods in large lecture classes. *Contemporary Educational Psychology*, *34*(1), 51–57. https://doi.org/10.1016/j.cedpsych.2008.04.002

McDaniel, M. A., Agarwal, P. K., Huelser, B. J., McDermott, K. B., & Roediger, H. L. (2011). Test-enhanced learning in a middle school science classroom: The effects of quiz frequency and placement. *Journal of Educational Psychology*, *103*(2), 399–414. https://doi.org/10.1037/a0021782

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

McDaniel, M. A., & Little, J. L. (2019). Multiple-choice and short-answer quizzing on equal footing in the classroom: Potential indirect effects of testing. In J. Dunlosky & K. A. Rawson (Eds.), *The Cambridge Handbook of Cognition and Education* (1st ed., pp. 480–499). Cambridge University Press. https://doi.org/10.1017/9781108235631.020

McDaniel, M. A., Thomas, R. C., Agarwal, P. K., McDermott, K. B., & Roediger, H. L. (2013). Quizzing in middle-school science: Successful transfer performance on classroom exams. *Applied Cognitive Psychology*, *27*(3), 360–372. https://doi.org/10.1002/acp.2914

McDaniel, M. A., Wildman, K. M., & Anderson, J. L. (2012). Using quizzes to enhance summative-assessment performance in a web-based class: An experimental study. *Journal of Applied Research in Memory and Cognition*, *1*(1), 18–26. https://doi.org/10.1016/j.jarmac.2011.10.001

McDermott, K. B., Agarwal, P. K., D'Antonio, L., Roediger, H. L., & McDaniel, M. A. (2014). Both multiple-choice and short-answer quizzes enhance later exam performance in middle and high school classes. *Journal of Experimental Psychology: Applied*, *20*(1), 3–21. https://doi.org/10.1037/xap0000004

Minear, M., Coane, J. H., Boland, S. C., Cooney, L. H., & Albat, M. (2018). The benefits of retrieval practice depend on item difficulty and intelligence. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. http://dx.doi.org/10.1037/xlm0000486

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*, *4:5*. https://doi.org/10/gf2rp4

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

Palmer, S., Chu, Y., & Persky, A. M. (2019). Comparison of rewatching class recordings versus retrieval practice as post-lecture learning strategies. *American Journal of Pharmaceutical Education*, *83*(9). https://doi.org/10.5688/ajpe7217

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/10.1037/0278-7393.31.1.3

Phelps, R. P. (2012). The effect of testing on student achievement, 1910–2010. *International Journal of Testing*, *12*(1), 21–43. https://doi.org/10.1080/15305058.2011.602920

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

Ramraje, S., & Sable, P. (2011). Comparison of the effect of post-instruction multiple-choice and short-answer tests on delayed retention learning. *The Australasian Medical Journal*, *4*, 332–339. https://doi.org/10.4066/AMJ.2011.727

Rawson, K. A., & Zamary, A. (2019). Why is free recall practice more effective than recognition practice for enhancing memory? Evaluating the relational processing hypothesis. *Journal of Memory and Language*, *105*, 141–152. https://doi.org/10/gfxx7n

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. https://doi.org/10.1111/j.1745-6916.2006.00012.x

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* (Vol. 55, pp. 1–36). Elsevier. 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), 1432–1463. https://doi.org/10.1037/a0037559

Rummer, R., Schweppe, J., & Schwede, A. (2019). Open-book versus closed-book tests in university classes: A field experiment. *Frontiers in Psychology*, *10:463*. https://doi.org/10/gfw84k

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

Stenlund, T., Jönsson, F. U., & Jonsson, B. (2017). Group discussions and test-enhanced learning: Individual learning outcomes and personality characteristics. *Educational Psychology*, *37*(2), 145–156. https://doi.org/10.1080/01443410.2016.1143087

Stenlund, T., Sundström, A., & Jonsson, B. (2016). Effects of repeated testing on short- and long-term memory performance across different test formats. *Educational Psychology*, *36*(10), 1710–1727. https://doi.org/10.1080/01443410.2014.953037

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

Vaughn, K. E., Rawson, K. A., & Pyc, M. A. (2013). Repeated retrieval practice and item difficulty: Does criterion learning eliminate item difficulty effects? *Psychonomic Bulletin & Review*, *20*(6), 1239–1245. https://doi.org/10.3758/s13423-013-0434-z

Vojdanoska, M., Cranney, J., & Newell, B. R. (2010). The testing effect: The role of feedback and collaboration in a tertiary classroom setting. *Applied Cognitive Psychology*, *24*(8), 1183–1195. https://doi.org/10.1002/acp.1630

Welch, S. (2019). An evaluation of Macmillan Education's LaunchPad as a textbook technology supplement when teaching introductory psychology. *Scholarship of Teaching and Learning in Psychology*, *5*(3), 236–246. https://doi.org/10.1037/stl0000162

Wheeler, M., Ewers, M., & Buonanno, J. (2003). Different rates of forgetting following study versus test trials. *Memory*, *11*(6), 571–580. https://doi.org/10.1080/09658210244000414

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

Chapter II

Examining the Testing Effect in University Teaching: Retrievability and Question Format Matter

Study 1

A version of this chapter is published as:

Greving, S., & Richter, T. (2018). Examining the testing effect in university teaching: Retrievability and question format matter. *Frontiers in Psychology*, *9:2412*. https://doi.org/10/gfkwvm

Examining the Testing Effect in University Teaching: Retrievability and Question Format Matter

Sven Greving & Tobias Richter

Abstract. Review of learned material is crucial for the learning process. One approach that promises to increase the effectiveness of reviewing during learning is to answer questions about the learning content rather than restudying the material (testing effect). This effect is well established in lab experiments. However, existing research in educational contexts has often combined testing with additional didactical measures that hampers the interpretation of testing effects. We aimed to examine the testing effect in its pure form by implementing a minimal intervention design in a university lecture (N = 92). The last 10 min of each lecture session were used for reviewing the lecture content by either answering short-answer questions, multiple-choice questions, or reading summarizing statements about core lecture content. Three unannounced criterial tests measured the retention of learning content at different times (1, 12, and 23 weeks after the last lecture). A positive testing effect emerged for short-answer questions that targeted information that participants could retrieve from memory. This effect was independent of the time of test. The results indicated no testing effect for multiple-choice testing. These results suggest that short-answer testing but not multiple-choice testing may benefit learning in higher education contexts.

Learners tend to remember less learning content when reading or listening to it only once (e.g., Aiken et al., 1975). Students often need to review the learned material, for example, when studying for exams. One potentially effective review strategy is the active retrieval of learned material from memory, which can be prompted by testing knowledge of the learned content. The finding that testing is superior to restudying the learning material is called the testing effect or retrieval practice effect (Roediger and Karpicke, 2006a). The superiority of testing compared to restudying might not be detected until later criterial tests or exams. Because of this latent effect, testing or retrieval practice is sometimes regarded as a desirable difficulty (Bjork, 1994). Desirable difficulties are defined as learning occasions that may hamper learning in the short run but enhance learning in the long run.

The testing effect is a robust finding in laboratory settings (e.g., Karpicke, 2017; Roediger and Karpicke, 2006b; Rowland, 2014), which has led researchers and practitioners to implement testing in applied educational contexts to promote the retention of learning content. Recent research has demonstrated the superiority of testing compared to restudying in various pedagogical settings (e.g., Karpicke, 2017, Table II.2). Based on these findings, several authors have advocated the use of tests in educational contexts to improve learning (Dunlosky et al., 2013; Dunlosky and Rawson, 2015; Dunn et al., 2013; McDaniel et al., 2007b).

Despite the promising results and recommendations, the generalizability to educational contexts and the conditions under which the effects occur remain an open question. Based on a review of key findings from lab experiments and a discussion of studies investigating the testing effect in real-world educational settings, we argue that many of the extant field studies suffer from limitations regarding the generalizability of the results. These limitations stem mostly from methodological problems such as a third variable that cofounds

the comparison of testing vs. restudying. In this article, we refer to the pure (unconfounded) difference between testing and restudying as the net testing effect. The aim of the present study was to examine the net testing effect in the real-world educational context of a university lecture.

The Testing Effect in Laboratory Settings

The testing effect has been a major focus of lab-based memory research for more than a century. Summarizing this research, recent meta-analyses by Adesope et al. (2017), Rowland (2014), and Phelps (2012) found a positive average testing effect with a medium to large effects size (Cohen's d/Hedges' g) ranging from 0.50 to 0.61. These meta-analyses also have identified moderators of the testing effect that are potentially relevant for applications in educational contexts.

Two factors that reliably affect the testing effect are feedback (Adesope et al., 2017; Rowland, 2014) and retrievability (Rowland, 2014). The provision of feedback, mostly in the form of presenting the correct answer, seems to increase the testing effect. Retrievability in this context describes the success with which learning content can be retrieved from memory, resulting in correct responses in the testing condition. Therefore, retrievability can be operationalized by the (reverse-scored) difficulty of items in the practice tests.

Conflicting results have been reported for different question formats used in the practice tests. In the meta-analysis by Adesope et al. (2017), multiple-choice questions elicited stronger testing effects than short-answer questions, whereas Rowland (2014) reported the opposite. Furthermore, a match between question format in the testing conditions and question format in the criterial tests seems to increase the testing effect according to the meta-analysis by Adesope et al., whereas this effect was not found by Rowland. In contrast to Adesope at al., Rowland excluded applied research in his meta-analysis. Therefore, the

divergent results of the two meta-analyses might reflect a moderating role of question format in educational contexts.

The Testing Effect in Educational Contexts

The robust testing effect found in laboratory experiments has spawned a growing body of research in educational contexts. One of the first studies of this kind was a study by McDaniel et al. (2007b). In this study, college students either took weekly quizzes in the form of short-answer questions or multiple-choice questions or they restudied previously learned content. Each condition was followed by feedback. In a later criterial test, short-answer testing led to a more pronounced testing effect than did multiple-choice testing.

Since then, the testing effect has been demonstrated in different age groups (for a review, see Dunlosky et al., 2013) and with learning materials of varying complexity (for a review, see Karpicke and Aue, 2015). Three meta-analyses (Adesope et al., 2017; Bangert-Drowns et al., 1991; Schwieren et al., 2017) reported a positive testing effect in educational contexts. Bangert-Drowns et al. included only research conducted in classrooms and reported a positive testing effect with an effect size of $d = 0.54$ for studies that compared testing and no testing. Adesope et al. analyzed all studies investigating the testing effect and included study setting (classroom vs. laboratory) as a moderator. This meta-analysis estimated a positive testing effect with an effect size of $g = 0.67$ for classroom settings. Finally, Schwieren et al. reported a positive testing effect of $d = 0.56$ for studies in which psychological learning content was taught in the classroom.

Although there seems to be a consensus among researchers that the testing effect occurs in real-world educational settings, little is known about factors that moderate the effect in such settings. Several studies have validated the moderating effects of feedback found in laboratory research in applied educational contexts (Downs, 2015; Marsh et al., 2012;

McDaniel et al., 2007b; Vojdanoska et al., 2010). Moreover, studies suggest that the testing effect can be found with different question formats in the practice tests (McDaniel et al., 2012; McDermott et al., 2014; Stenlund et al., 2016). The match between question formats in testing and criterial tests does not seem to matter (McDermott et al., 2014).

Limitations of Previous Research on the Testing Effect in Educational Contexts

Numerous studies have investigated the testing effect in real-world educational contexts. However, many of these studies provide only limited information on the current research question because of internal or external validity problems that hamper the interpretation of the results.

One limiting feature of many extant studies on the testing effect in applied contexts is a lack of randomization. Because of practical constraints, researchers have often employed a quasi-experimental design, for example, by varying independent variables between courses, sections, or years (Batsell et al., 2017; Cranney et al., 2009; Khanna, 2015; Leeming, 2002; Mayer et al., 2009; Vojdanoska et al., 2010). The internal validity of these studies is questionable, because the extent that differences between the testing and the control condition are attributable to other (uncontrolled) differences between the groups is uncertain.

Other studies are limited because they lack a restudy control condition but compare the testing condition to conditions in which no exposure to information subsequent to the initial learning took place (Batsell et al., 2017; Bell et al., 2015; Downs, 2015; Foss and Pirozzolo, 2017; Johnson and Kiviniemi, 2009; Khanna, 2015; Lyle and Crawford, 2011; Marsh et al., 2012; Mayer et al., 2009; McDaniel et al., 2007a, 2011, 2013; Roediger et al., 2011; Shapiro and Gordon, 2012; Vojdanoska et al., 2010). In these studies, the testing effect is confounded with differences in exposure to and engagement with learning content, which severely limits the interpretation of their findings. To assess the magnitude of the testing effect in applied educational settings, comparing testing conditions with restudy conditions or

other activities that are assumed to promote the retention of information is essential (for examples, see Adesope et al., 2017; Rummer et al., 2017).

A third limitation threatening the internal validity is found in studies that allow participants to repeat tests on the same subject. Some studies limit the amount of repetitions (Wiklund-Hörnqvist et al., 2014) while others do not (Bell et al., 2015; Downs, 2015; Johnson and Kiviniemi, 2009; McDaniel et al., 2012; Yong and Lim, 2016). Even when participants are also free to restudy the material as often as they like, it remains unclear whether differences in learning outcomes are solely attributable to testing vs. no testing or whether additional factors (e.g., differential effects of motivation) influence the number of repetitions and thus the learning outcomes.

A fourth limitation is that many studies combine the testing conditions with feedback (Bell et al., 2015; Carpenter et al., 2009; Cranney et al., 2009; Downs, 2015; Leeming, 2002; Lyle and Crawford, 2011; McDaniel et al., 2007a, 2007b, 2011, 2012; Stenlund et al., 2017; Wiklund-Hörnqvist et al., 2014). Research has shown that testing may profit from feedback in educational settings (Vojdanoska et al., 2010). However, feedback also provides an additional study opportunity and thus an additional exposure to the learning content. We therefore argue that effects obtained in studies that combined testing with feedback cannot be readily interpreted in terms of a testing effect.

A fifth limitation is present in so-called open-label studies (Batsell et al., 2017; Bing, 1984; Daniel and Broida, 2004). In such studies, participants are told beforehand whether the learning content is tested or not, which might alter learning behavior and strategies between conditions when learning (Finley and Benjamin, 2012). As a consequence, differences obtained in testing vs. no-testing conditions can be due to differences in learning

behavior that learners in the testing condition engage in, because they anticipate learning content. That is, the differences might not be due to the testing effect.

Furthermore, the internal validity is threatened in studies that feature high-stakes testing conditions (Batsell et al., 2017; Leeming, 2002; Lyle and Crawford, 2011). In these studies, participants' scores in the testing condition affect the participants' grades. This fact hampers the interpretation of testing effects in two ways. First, unannounced high-stakes tests have been shown to reduce the benefit of testing in applied educational settings compared to unannounced low-stakes tests (Khanna, 2015). Second, whenever open-label studies also include high-stakes testing conditions, students might alter their learning behavior and strategies, because they are motivated to get good grades.

Finally, some researchers have opted to avoid the difficulties associated with implementing experimental designs in real-world educational settings by conducting lab-based studies with "educationally relevant materials" (Butler and Roediger, 2007; Einstein et al., 2012; Marsh et al., 2012; Stenlund et al., 2016; Yong and Lim, 2016). This approach neglects the problem that the learning in secondary or postsecondary courses is likely to differ in terms of motivation, personal involvement, and effort from learning only for the purpose of participating in a psychological or educational study. These differences pose a threat to the external validity of such studies and limit their generalizability to the testing effect in actual educational settings.

Theoretical Framework and Rationale of the Present Study

The aim of the present study was to examine the testing effect in an authentic educational setting of a university lecture with an experimental design that minimizes the issues that limit the validity of previous field studies. We used an experimental design that compared testing on a single occasion without the provision of feedback with a restudy

condition. Furthermore, participants' results in the testing conditions would not affect their grades and participants would not know the type of review condition to expect after learning.

Investigating the testing effect in this fashion is informative for a number of reasons. First, most field experiments to date include features that limit the interpretation of the results. In order to investigate the net testing effect in educational contexts, we excluded all features that might cloud the interpretability of this effect. Furthermore, in real world educational contexts, it is not always possible to provide feedback during practice tests or to provide multiple opportunities to practice retrieval. Furthermore, a single opportunity to practice retrieval without feedback makes low demands on time and personal resources compared to multiple retrieval practice opportunities with feedback. Investigating whether testing on a single occasion without feedback is effective can thus be relevant for future research and practitioners alike.

Most theories of the testing effect assume that even in this minimalistic setting, retrieval would be more beneficial for retention than restudying. The desirable difficulty framework (Bjork, 1994), the new theory of disuse (e.g., Bjork & Bjork, 2011), and the retrieval effort hypothesis (Pyc & Rawson, 2009) all incorporate the assumption that effortful retrieval should lead to better retention of that learning content and thus testing should lead to better retention than does restudying. However, it should be noted that in all of these theoretical notions retrievability plays a crucial role. Whenever the correct information cannot be retrieved from memory, no beneficial effects compared to restudying may be expected (e.g., Jang et al., 2012).

It has been repeatedly argued that multiple-choice questions and short-answer questions differ in the effort needed to be answered correctly and—given these theoretical underpinnings—should consequently lead to different testing effects (e.g., Karpicke, 2017).

These different testing effects have already been demonstrated in educational contexts (McDaniel et al., 2007b).

Researchers and practitioners do not always use verbatim repetitions of retrieval practice in criterial tests and exams. Instead, questions are used that ask for related information. Previous studies suggest that these questions may lead to impaired retrieval—a phenomenon dubbed retrieval induced forgetting (for an overview, see Bjork et al., 2014)—and that this impairment depends on the question format (Carroll et al., 2007). Furthermore, research has also demonstrated that retrieval practice promoted retention of learning content not subject to retrieval practice (for an overview, see Pan and Rickard, 2018) and that the design of multiple-choice questions may affect whether unrelated learning content benefits from retrieval practice (Little et al., 2012). To investigate the potential moderating role of question format, we implemented two different testing conditions, one with short-answer questions and the other with multiple-choice questions in the practice test.

The experiment was conducted in a university lecture with minimal intervention. Therefore, the learning content was the regular course material and the lecture was held as usual. The intervention took place in the last 10 min of a 90-min lesson. Furthermore, we measured learning outcomes (i.e., memory for the learning content) in criterial tests at three different times: before and after the final exam and half a year after the final exam. In the criterial tests, we also included questions that were not targeted in the testing conditions but contained related information as well as questions that targeted learning content not subject to testing or restudy, in order to control for differential effects of these question types on multiple-choice and short-answer testing.

We expected a positive testing effect to occur. Furthermore, we examined as exploratory research questions whether the testing effect would depend on question format in the practice tests, the time of the criterial test, and retrievability. We reasoned that short-

answer questions would be more suitable for prompting active retrieval of knowledge, leading to a stronger testing effect. Moreover, assuming that testing is a desirable learning difficulty, the benefits of testing vs. restudying might become visible, particularly at later criterial tests. Finally, retrievability might matter because the testing effect can only occur when retrieval is successful, especially when no feedback is given for responses in the practice tests.

Method

Participants

Participants were 137 undergraduate students in their first semester, most of them female (71%) and students of psychology (92%). They participated in at least one lecture session and one criterial test. All students gave their informed and written consent prior to participation. Participants' age ranged between 18 and 74 with a mean age of 23.15 (*SD* = 7.74).

Materials

Test questions and restudy statements

The content of seven lecture sessions of an introductory lecture in cognitive psychology was surveyed and 24 information units per session were identified. For each information unit, one summarizing statement, one short-answer question and one multiple-choice question were created. Statements were created by summarizing the key information of the information unit in one sentence (e.g., "Prosopagnosia is a cognitive disorder of face perception in which the ability to recognize faces is impaired to the extent that the person becomes blind to faces."). Short-answer questions were created by asking for the key information of the information unit (e.g. "What is prosopagnosia?"). Multiple-choice questions were created by adding four response options with only one correct answer to the

short-answer question (e.g., “What is prosopagnosia? (A) face blindness, (B) shape blindness, (C) color blindness, (D) object blindness”).

Revision materials

For each of the seven lecture sessions, eight information units were randomly drawn from the 24 information units prepared for this session. Based on the selected information units, revision materials were prepared for each lecture session. The revision materials consisted of a one-page questionnaire asking for basic demographic information and two pages of revision items corresponding to the selected information units, consisting of either (a) eight summarizing statements (restudy condition), (b) eight short-answer questions (testing, short-answer questions), or (c) eight multiple-choice questions (testing, multiple-choice condition). In all three versions, information units were presented in the same order with four information units on each page.

Criterial tests

Three criterial tests (Criterial Tests 1 to 3) were constructed that consisted of questions based on the pool of 24 information units determined for each of the seven lecture sessions. The pool of questions was expanded by creating alternate versions of the questions used in the revision material. Alternate questions were created by asking for the key information in another way (e.g., “What is the medical term for face blindness?”). For each information unit, an alternate short-answer question and an alternate multiple-choice question were created.

Each of the three criterial tests consisted of three components: (a) questions corresponding to information units included in the revision materials, (b) questions corresponding to information units not included in the revision materials, and (c) alternate questions, corresponding to information units but not identical to questions included in the

revision materials. Additionally, questions previously asked in criterial tests were also included in Criterial Tests 2 and 3. Table II.1 depicts the composition of the criterial tests and the total number of questions per criterial test. Most notably, the composition of Criterial Test 3 differed from the composition of the other two criterial tests. This difference was due to a sampling error in the composition of the criterial tests.

Table II.1

Criterial Test Composition by Components and Repetition of Questions in Later Criterial Tests

	Criterial Test 1			Criterial Test 2			Criterial Test 3		
	Questions Included in Study Material			Questions Included in Study Material			Questions Included in Study Material		
Previously Tested in Criterial Test	Verbatim	Alternate	New Questions	Verbatim	Alternate	New Questions	Verbatim	Alternate	New Questions
Yes				7	7	7	7	7	7
No	14	14	14	14	14	14	0[a]	0[a]	7
Total		42			63			28	

Note. [a]Not included because of a sampling error in the composition of the criterial tests.

Each criterial test consisted of short-answer questions and multiple-choice questions in equal proportions. Two versions were created (Versions A and B) by altering the order of questions and the question format (i.e., multiple-choice questions vs. short-answer questions) of the same question between criterial test versions so that all multiple-choice questions in Version A were short-answer questions in Version B and vice versa. All study materials are made available upon request to interested researchers.

Scoring

Multiple-choice questions were scored with 1 when only the correct option was ticked (correct answer) vs. 0 when a distractor was ticked or no response was given (incorrect or missing response). Short-answer questions were scored with 1 (correct response) vs. 0 (incorrect or missing response). Two independent raters scored all responses to short-answer questions. Inter-rater reliability was high across all lectures and criterial tests (6855 observations, Cohen's $\kappa = .87$) and thus scores from only one rater was included in the analyses. The performance scores based on both question types served as dependent variable.

Procedure

General procedure

The study was conducted over a period of two semesters. In the first semester (October 2015–February 2016), a weekly introductory psychology lecture was taught that covered basic principles of cognitive psychology. In lecture Sessions 4 to 10, the manipulation of review condition (testing with multiple-choice or short-answer questions) took place. The three criterial tests, which assessed the learning of content taught in the seven lecture sessions, were administered unannounced to the students at scheduled times after the last lecture with learning content (i.e., after Session 10). Criterial Test 1 was administered one week after Session 10. Criterial Test 2 was administered in the first session of the second

semester (April 2016–July 2016), 12 weeks after Session 10, and Criterial Test 3 was administered in the final session of the second semester, 23 weeks after Session 10.

Procedure during the lecture sessions

In each of the lecture Sessions 4–10, the last 10 min were reserved for the manipulation of the review condition. Participation was voluntary. Students were allowed to leave the lecture hall after the end of the regular lecture. Research assistants then administered the review materials, assigning participants randomly to one of the three review conditions (testing with multiple-choice questions, testing with short-answer questions, or restudy). Participants first filled in basic demographic information. They were then given 4 min to complete each page of the two pages of revision items. This was the sole opportunity to review the learning content according to one of the three conditions. Finally, participants were thanked for their participation, and the materials were collected.

Criterial tests

All students present in the respective lecture sessions were allowed to take Criterial Test 1, 2, or 3, irrespective of previous participation in the study. In each of these sessions, the two versions of the criterial test were then administered in an alternating way so that participants sitting next to each other received different versions. Students were allowed 45 min to complete the test and could leave when they finished.

Design

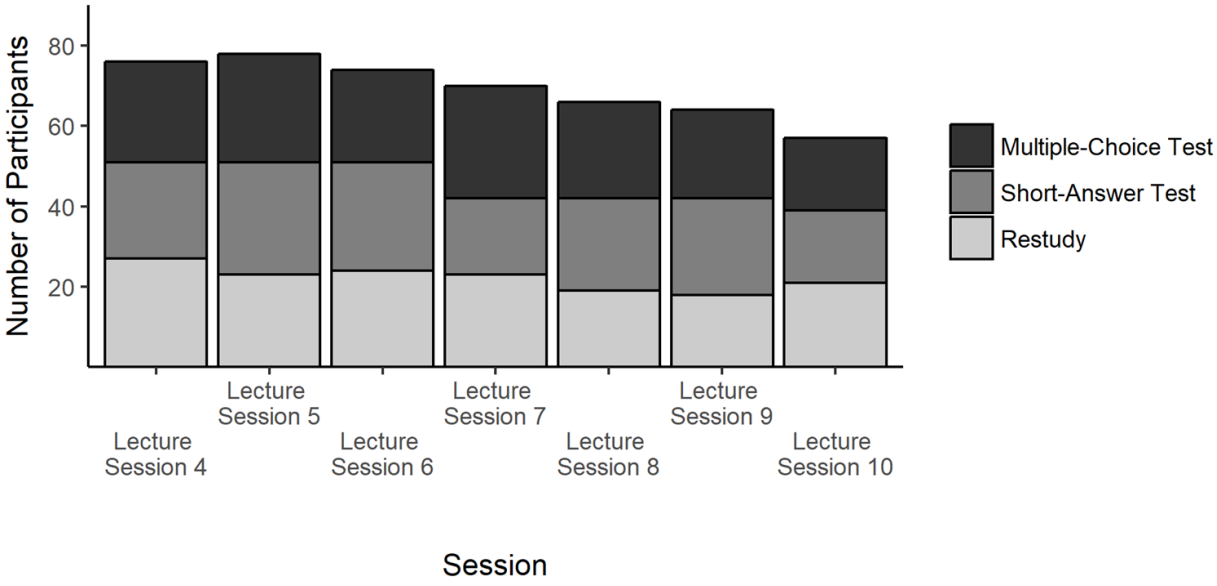
The design was a 3 x 3 within-subjects design with the independent variables review condition (multiple-choice test, short-answer test, restudy) and time of test (Criterial Tests 1–3 at 1, 12, and 23 weeks after the final lecture session). Each participant received one of two versions of each criterial test, which differed in format (short-answer vs. multiple-choice

question) and order of questions. The dependent variable was the performance (percent correct) on the multiple-choice and short-answer questions in the criterial tests.

The design was implemented by randomly assigning participants in Sessions 4–10 of the focal lecture to one of the three review conditions. Likewise, participants were assigned to one of the two test versions of the criterial tests administered at each time of test. Figure II.1 depicts the number of participants that were assigned to each review condition in the seven lecture sessions. The random allocation led to equal distributions of participants across review conditions. Similarly, participants were evenly distributed to the criterial test versions (Versions A:B) in Criterial Tests 1 (n = 32:33), 2 (n = 40:40), and 3 (n = 25:28). We assume that missing data is missing completely at random and thus inferences can proceed by analyzing the observed data only (Ibrahim and Molenberghs, 2009).

Figure II.1

Participation by Condition and Session



Results

We estimated generalized linear mixed effect models (GLMMs) with a logit-link function (Dixon, 2008) with the R package lme4 (Bates et al., 2015).

For comparisons between conditions and extracting mean performance scores for different experimental conditions the R package lsmeans was used (Lenth, 2016). For all significance tests, Type I error probability was set to .05 (one-tailed for testing directed hypotheses). Participants and test items were included as random effects (random intercepts) in all models.

Separate models were estimated to examine the testing effect based on short-answer questions and the testing effect based on multiple-choice questions. In each of the two models, the testing condition was compared to the restudy condition that involved reading the summarizing statements that provided the correct answer (dummy-coded: testing = 1, restudy = 0). We additionally tested whether the testing effect depended on the time of the criterial

test by including two dummy-coded predictors for Criterial Test 2 and Criterial Test 3 (Criterial Test 1 was the reference condition coded with 0 in both predictors) and the interactions of these predictors with testing vs. restudying. In addition, the models included the retrievability of learned information in form of two dummy-coded predictors that contrasted items of medium retrievability and low retrievability with items of high retrievability as the reference condition. We examined whether higher retrievability rates were associated with a larger testing effect. To construct this predictor, we grouped the short-answer questions and the multiple-choice questions separately into three equally sized, ordered categories (tertiles) according to their difficulty in the practice tests. To avoid distortions from extreme values, we discarded the lowest and the highest 5% of the distribution before the grouping. Item difficulties to the multiple-choice questions were corrected for guessing. For each of the two item types (short-answer and multiple-choice questions), grouping resulted in three categories of items with high (short-answer questions: item difficulties from 46% to 81%; multiple-choice questions: 78% to 100%), medium (short answer questions: 25% to 45%; multiple-choice questions: 53% to 77%), or low retrievability (short answer questions: 5% to 24%; multiple-choice questions: 0% to 53%). Finally, the models included the interaction of retrievability with testing vs. restudying. All predictors and their interactions were entered simultaneously in the models.

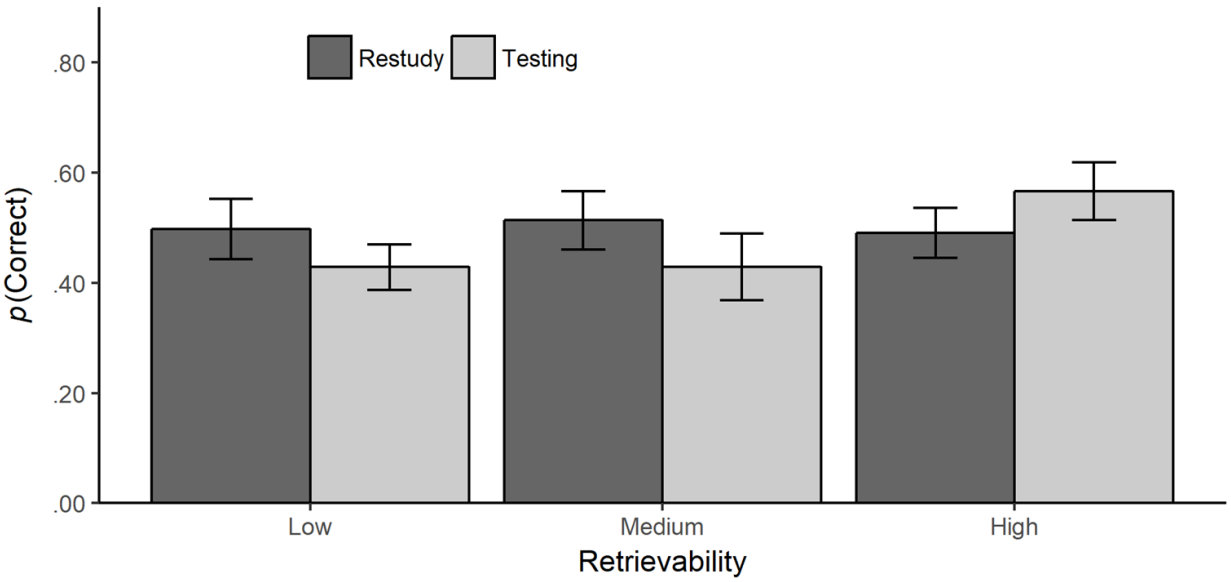
Effects of Testing with Short-Answer Questions

The model estimates for the effects of testing with short-answer questions can be found in Table II.2 (left columns). This model revealed a positive effect for testing ($\beta = 0.44$, $SE = 0.24$, $p = .033$, one-tailed). However, the interaction of testing vs. restudying with the predictor comparing low to high retrievability was significant ($\beta = -0.60$, $SE = 0.28$, $p = .016$, one-tailed). Likewise, the interaction with the predictor comparing medium to high

retrievability was significant (β = -0.66, *SE* = 0.35, *p* = .030, one-tailed). Planned contrasts revealed a testing effect only for items with high retrievability (*z* = 1.85, *p* = .032, one-tailed) but not for items with medium (*z* = -0.74, *p* = .771, one-tailed) or low retrievability (*z* = -0.66, *p* = .746, one-tailed) (Figure II.2).

Figure II.2

Testing with short-answer questions



Note. Mean probability of correct responses (with standard errors) in all criterial test items (back-transformed from the logits in the GLMM) by retrievability and review condition (testing vs. restudy).

Table II.2

Model Parameters

	Short-Answer Questions				Multiple-Choice Questions			
Parameter	β	*SE*	*z*	*p*	β	*SE*	*z*	*p*
Intercept	-0.34	0.25	-1.36	.173	0.07	0.29	0.25	.803
Testing	0.44	0.24	1.84	.033[a]	-0.42	0.24	-1.76	.078
Criterial Test 2	0.80	0.27	2.96	.003	0.74	0.29	2.55	.011
Criterial Test 3	0.11	0.38	0.30	.768	0.38	0.42	0.91	.361
Testing x Criterial Test 2	-0.14	0.22	-0.63	.531	0.12	0.22	0.55	.583
Testing x Criterial Test 3	-0.23	0.32	-0.73	.468	-0.11	0.34	-0.33	.739
Low retrievability	0.03	0.25	0.10	.917	-0.31	0.25	-1.23	.219
Medium retrievability	0.09	0.23	0.40	.692	-0.35	0.27	-1.33	.184
Testing x Low retrievability	-0.60	0.28	-2.14	.016[a]	0.17	0.27	0.62	.534
Testing x Medium retrievability	-0.66	0.35	-1.88	.030[a]	0.060	0.37	0.16	.872
$N_{Participants}$	92				91			
N_{Items}	77				77			

Note. Parameter estimates for the models estimating the effect of testing with short-answer questions and multiple-choice questions, time of test, and retrievability on short-answer questions and multiple-choice questions on learning performance in the criterial tests. Testing (dummy-coded: testing = 1, restudy = 0). Criterial Test 2 (dummy-coded: Criterial Test 2 = 1, Criterial Test 1 = 0). Criterial Test 3 (dummy-coded: Criterial Test 3 = 1, Criterial Test 1 = 0). Low retrievability (dummy-coded: low retrievability = 1, high retrievability = 0). Medium retrievability (dummy-coded: medium retrievability = 1, high retrievability = 0). [a]*p* values refer to one-tailed tests for $\beta > 0$. Other *p* values refer to two-tailed tests.

The interactions with time of tests were not significant, suggesting that the testing effect obtained for short-answer questions was independent of the time of test. However, there was a main effect of the predictor comparing Criterial Test 2 to Criterial Test 1. The probability of giving a correct response was higher at Criterial Test 2 (P = .61, SE = .04) compared to Criterial Test 1 (P = .43, SE = .05).

Effects of Testing with Multiple-Choice Questions

The model estimates for the effects of testing with multiple-choice questions can be found in Table II.2 (right columns). No effect of testing vs. restudying emerged. None of the interaction effects of testing with time of test or retrievability were significant. Again, there was a main effect of the predictor comparing Criterial Test 2 to Criterial Test 1. The probability of correct responses was higher at Criterial Test 2 (*P* = .62, *SE* = .05) compared to Criterial Test 1 (*P* = .42, *SE* = .05).

In sum, the results indicated no testing effect for multiple-choice questions.

Discussion

The present study investigated the testing effect in a university education setting by implementing a minimal intervention in an existing university course. In contrast to many previous studies with a similar aim, we took care to avoid confounding factors and based our study on an experimental design. The main finding was a testing effect for practice tests based on short-answer questions, provided that participants in the testing condition were able to retrieve this content. No evidence was found for a testing effect for practice tests based on multiple-choice questions.

Our study method shares many features with lab experiments investigating the net testing effect (e.g., Roediger and Karpicke, 2006a, Experiment 1), with the obvious difference being that the setting of the current experiment was in real-world educational

context. Although this difference alone could have contributed to the lack of an overall testing effect, two other factors are likely to affect the testing effect in laboratory and educational contexts. Most research uses a repetition of the entire learning content in the restudy condition, but exact repetitions are difficult to implement in real-world educational contexts because of time constraints, that is, usually only selected information is restudied. Participants in our study studied summaries of important aspects of the lecture. In this regard, Kornell et al. (2012) argued that restudying the material in the same way might overestimate the testing effect, but they also provided evidence that testing might be superior to restudying non-exact repetition of study material.

The testing effect for practice tests based on short answer questions depended on retrievability of the initially learned content. A testing effect occurred for only questions with a high retrievability, that is, mean retrievability rates between 46% to 81%. This finding is in line with previous findings from laboratory experiments (Rowland, 2014) and with the bifurcation model (Halamish and Bjork, 2011; Kornell et al., 2011). The bifurcation model states that the superiority of testing without feedback compared to restudying depends on the amount of successfully retrieved items in the testing condition. Support for the bifurcation model comes from Rowlands (2014) meta-analysis that revealed no testing effect for laboratory experiments with no corrective feedback and retrievability rates of less than or equal to 50%. Our findings can thus be regarded as additional support of the bifurcation model in educational contexts. These findings also extend the existing research, because the testing effect, although implemented through a minimalistic intervention, was stable over a period of at least 23 weeks.

In line with findings from lab experiments investigating the net testing effect, a testing effect emerged for short-answer questions after a single presentation of these

questions. Lab experiments investigating repeated testing without feedback also revealed a net testing effect (Roediger and Karpicke, 2006a, Experiment 2; Wirebring et al., 2015). Repeated short-answer testing might be even more potent in an educational setting than short-answer testing on a single occasion. Future studies should compare these two ways to implement short-answer testing in educational settings.

In contrast to testing based on short-answer questions, no testing effect emerged for practice tests based on multiple-choice questions. This pattern of effects is in line with current theories of the testing effect that emphasize the role of cognitive effort during retrieval (Bjork, 1994; Pyc and Rawson, 2009). Questions that prompt effortful retrieval are likely to elicit stronger testing effects. The multiple-choice questions used in the present study were relatively easy (compared to the short-answer questions). Two-thirds of the items were solved correctly in most of the cases, suggesting that participants spent relatively little effort in retrieving the relevant information from long-term memory. Moreover, multiple-choice questions may have a negative effect on learning retention because of the presence of distracters (lures). Roediger and Marsh (2005) have shown that multiple-choice testing may lead participants to answer later criterial tests with false information. Further research suggests that this impact can be lessened by corrective feedback (Marsh et al., 2012). In the present study, no corrective feedback was given, implying that the distracting information could have influenced the performance on the criterial tests, counteracting the testing effect.

The experimental design in a field study is a strength of the present study, but the method also presents some limitations. Compared to laboratory experiments, external influences potentially play a much greater role in a field setting. For the present study, the extent that other factors (e.g., metamemorial, metacognitive, or motivational factors) influenced learning behavior during lectures and review conditions, when taking the criterial tests, or in the days and weeks between the lectures and the criterial tests is unknown. For

example, the performance in the criterial tests increased steeply from the first to the second criterial test, which is likely caused by participants' increased study activities in preparation for the upcoming exam. Participation in the study in each of the lectures was voluntary, which might have caused selection effects. However, it must be noted that these selection effects likely affected all experimental conditions to the same extent, because participants were unaware of the review condition that they would be assigned to when they made their decision to participate.

Another limitation that our study shares with other studies on the testing effect is the potential confound of test properties for the practice and criterial tests. For example, multiple-choice questions not eliciting a testing effect might be due to the low demand on retrieval effort involved in answering multiple-choice questions (e.g., Nguyen and McDaniel, 2015). Thus, drawing conclusions that multiple-choice questions are generally unsuitable for eliciting a testing effect would be premature.

To conclude, this research contributes to the literature by demonstrating a testing effect for practice tests with short-answer questions in the real-world educational context of a university lecture. Previous research has examined the testing effect, normally combined with additional features or based on quasi-experimental designs, which has hindered interpretation of the testing effect reported in these studies. In contrast, the present study provides clear evidence for the claim that answering short answer questions only once and without feedback, compared to restudying key points of the lecture, benefits retention of learning content even beyond the final exam. However, one important condition is that the difficulty of these questions must be at a level such that students are able to answer most of these questions correctly. To use the testing effect to foster learning, educational practitioners should identify the most important topics of their lecture, teach these thoroughly, and use

short-answer testing to solidify the knowledge about these topics. Finally, presenting students with multiple-choice questions might be ineffective, compared to restudying key points of the lecture. Given these findings, we advise practitioners to use short-answer testing rather than multiple-choice testing to foster learning in university lectures.

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

Aiken, E. G., Thomas, G. S., & Shennum, W. A. (1975). Memory for a lecture: Effects of notes, lecture rate, and informational density. *Journal of Educational Psychology*, *67*(3), 439–444. https://doi.org/10.1037/h0076613

Bangert-Drowns, R. L., Kulik, J. A., & Kulik, C.-L. C. (1991). Effects of frequent classroom testing. *The Journal of Educational Research*, *85*(2), 89–99. https://doi.org/10.1080/00220671.1991.10702818

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01

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

Bing, S. B. (1984). Effects of testing versus review on rote and conceptual learning from prose. *Instructional Science*, *13*(2), 193–198. https://doi.org/10.1007/BF00052385

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). MIT Press.

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., Pashler, H., & Cepeda, N. J. (2009). Using tests to enhance 8th grade students' retention of U.S. history facts. *Applied Cognitive Psychology*, *23*(6), 760–771. https://doi.org/10.1002/acp.1507

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

Daniel, D. B., & Broida, J. (2004). Using web-based quizzing to improve exam performance: Lessons learned. *Teaching of Psychology*, *31*(3), 207–208. https://doi.org/10.1207/s15328023top3103_6

Dixon, P. (2008). Models of accuracy in repeated-measures designs. *Journal of Memory and Language*, *59*(4), 447–456. https://doi.org/10.1016/j.jml.2007.11.004

Downs, S. D. (2015). Testing in the college classroom: Do testing and feedback influence grades throughout an entire semester? *Scholarship of Teaching and Learning in Psychology*, *1*(2), 172–181. https://doi.org/10.1037/stl0000025

Dunlosky, J., & Rawson, K. A. (2015). Practice tests, spaced practice, and successive relearning: Tips for classroom use and for guiding students' learning. *Scholarship of Teaching and Learning in Psychology*, *1*(1), 72–78. https://doi.org/10.1037/stl0000024

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

Dunn, D. S., Saville, B. K., Baker, S. C., & Marek, P. (2013). Evidence-based teaching: Tools and techniques that promote learning in the psychology classroom. *Australian Journal of Psychology*, *65*(1), 5–13. https://doi.org/10.1111/ajpy.12004

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

Finley, J. R., & Benjamin, A. S. (2012). Adaptive and qualitative changes in encoding strategy with experience: Evidence from the test-expectancy paradigm. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *38*(3), 632–652. https://doi.org/10.1037/a0026215

Foss, D. J., & Pirozzolo, J. W. (2017). Four semesters investigating frequency of testing, the testing effect, and transfer of training. *Journal of Educational Psychology*. https://doi.org/10.1037/edu0000197

Halamish, V., & Bjork, R. A. (2011). When does testing enhance retention? A distribution-based interpretation of retrieval as a memory modifier. *Journal of Experimental Psychology. Learning, Memory & Cognition*, *37*(4), 801–812. https://doi.org/10.1037/a0023219

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

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

Karpicke, J. D., & Aue, W. R. (2015). The testing effect is alive and well with complex materials. *Educational Psychology Review*, *27*(2), 317–326. https://doi.org/10.1007/s10648-015-9309-3

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

Kornell, N., Bjork, R. A., & Garcia, M. A. (2011). Why tests appear to prevent forgetting: A distribution-based bifurcation model. *Journal of Memory and Language*, *65*(2), 85–97. https://doi.org/10.1016/j.jml.2011.04.002

Kornell, N., Rabelo, V. C., & Klein, P. J. (2012). Tests enhance learning—Compared to what? *Journal of Applied Research in Memory and Cognition*, *1*(4), 257–259. https://doi.org/10.1016/j.jarmac.2012.10.002

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

Lenth, R. V. (2016). Least-squares means: The R package lsmeans. *Journal of Statistical Software*, *69*(1). https://doi.org/10.18637/jss.v069.i01

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

Marsh, E. J., Fazio, L. K., & Goswick, A. E. (2012). Memorial consequences of testing school-aged children. *Memory*, *20*(8), 899–906. https://doi.org/10.1080/09658211.2012.708757

Mayer, R. E., Stull, A., DeLeeuw, K., Almeroth, K., Bimber, B., Chun, D., Bulger, M., Campbell, J., Knight, A., & Zhang, H. (2009). Clickers in college classrooms: Fostering

learning with questioning methods in large lecture classes. *Contemporary Educational Psychology*, *34*(1), 51–57. https://doi.org/10.1016/j.cedpsych.2008.04.002

McDaniel, M. A., Agarwal, P. K., Huelser, B. J., McDermott, K. B., & Roediger, H. L. (2011). Test-enhanced learning in a middle school science classroom: The effects of quiz frequency and placement. *Journal of Educational Psychology*, *103*(2), 399–414. https://doi.org/10.1037/a0021782

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

McDaniel, M. A., Roediger, H. L., & McDermott, K. B. (2007). Generalizing test-enhanced learning from the laboratory to the classroom. *Psychonomic Bulletin & Review*, *14*(2), 200–206. https://doi.org/10.3758/BF03194052

McDaniel, M. A., Thomas, R. C., Agarwal, P. K., McDermott, K. B., & Roediger, H. L. (2013). Quizzing in middle-school science: Successful transfer performance on classroom exams. *Applied Cognitive Psychology*, *27*(3), 360–372. https://doi.org/10.1002/acp.2914

McDaniel, M. A., Wildman, K. M., & Anderson, J. L. (2012). Using quizzes to enhance summative-assessment performance in a web-based class: An experimental study. *Journal of Applied Research in Memory and Cognition*, *1*(1), 18–26. https://doi.org/10.1016/j.jarmac.2011.10.001

McDermott, K. B., Agarwal, P. K., D'Antonio, L., Roediger, H. L., & McDaniel, M. A. (2014). Both multiple-choice and short-answer quizzes enhance later exam performance in middle and high school classes. *Journal of Experimental Psychology: Applied*, *20*(1), 3–21. https://doi.org/10.1037/xap0000004

Nguyen, K., & McDaniel, M. A. (2015). Using quizzing to assist student learning in the classroom: The good, the bad, and the ugly. *Teaching of Psychology*, *42*(1), 87–92. https://doi.org/10.1177/0098628314562685

Phelps, R. P. (2012). The effect of testing on student achievement, 1910–2010. *International Journal of Testing*, *12*(1), 21–43. https://doi.org/10.1080/15305058.2011.602920

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

Roediger, H. L., Agarwal, P. K., McDaniel, M. A., & McDermott, K. B. (2011). Test-enhanced learning in the classroom: Long-term improvements from quizzing. *Journal of Experimental Psychology: Applied*, *17*(4), 382–395. https://doi.org/10.1037/a0026252

Roediger, H. L., & Karpicke, J. D. (2006a). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*(3), 181–210. https://doi.org/10.1111/j.1745-6916.2006.00012.x

Roediger, H. L., & Karpicke, J. D. (2006b). 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., & Marsh, E. J. (2005). The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(5), 1155–1159. https://doi.org/10.1037/0278-7393.31.5.1155

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin*, *140*(6), 1432–1463. https://doi.org/10.1037/a0037559

Rummer, R., Schweppe, J., Gerst, K., & Wagner, S. (2017). Is testing a more effective learning strategy than note-taking? *Journal of Experimental Psychology: Applied*, *23*(3), 293–300. https://doi.org/10.1037/xap0000134

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

Stenlund, T., Jönsson, F. U., & Jonsson, B. (2017). Group discussions and test-enhanced learning: Individual learning outcomes and personality characteristics. *Educational Psychology*, *37*(2), 145–156. https://doi.org/10.1080/01443410.2016.1143087

Stenlund, T., Sundström, A., & Jonsson, B. (2016). Effects of repeated testing on short- and long-term memory performance across different test formats. *Educational Psychology*, *36*(10), 1710–1727. https://doi.org/10.1080/01443410.2014.953037

Vojdanoska, M., Cranney, J., & Newell, B. R. (2010). The testing effect: The role of feedback and collaboration in a tertiary classroom setting. *Applied Cognitive Psychology*, *24*(8), 1183–1195. https://doi.org/10.1002/acp.1630

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

Yong, P. Z., & Lim, S. W. H. (2016). Observing the testing effect using coursera video-recorded lectures: A preliminary study. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.02064

Chapter III

Practicing Retrieval in University Teaching: Short-Answer Questions Are Beneficial, Whereas Multiple-Choice Questions Are Not

Study 2

A version of this chapter was published as:

Greving, S., & Richter, T. (2022). Practicing retrieval in university teaching: Short-answer questions are beneficial, whereas multiple-choice questions are not. *Journal of Cognitive Psychology* (online first). https://doi.org/10.1080/20445911.2022.2085281

Practicing Retrieval in University Teaching: Short-Answer Questions Are Beneficial, Whereas Multiple-Choice Questions Are Not

Sven Greving & Tobias Richter

Abstract. Proponents of the testing effect claim that answering questions from memory about the learning content benefits retention of learning content more than does additional restudying—even when learners do not receive feedback in form of correct answers. In educational contexts, evidence for this claim is scarce and points toward differential effects for different question formats: Benefits emerged for short-answer questions but not for multiple-choice questions, presumably because multiple-choice questions require cognitive effort and active retrieval to a lesser extent than short-answer questions. The present study implemented a minimal intervention design in five sessions of an introductory psychology lecture. In each session, participants reviewed core lecture content by answering short-answer questions, multiple-choice questions, or reading summarizing statements. An unannounced test measured the retention of learning content. Bayesian analyses revealed a positive testing effect for short-answer questions that was strongest in items that were most difficult to retrieve. Analyses also provided evidence for the absence of a testing effect for multiple-choice questions. These results suggest that short-answer testing but not multiple-choice testing is beneficial in higher education contexts, even without feedback.

Chapter III

Practitioners need to know what techniques work when trying to foster retention of learned content. Research on the testing effect suggests that taking tests is more effective for retention than other practice techniques such as restudying (e.g., Karpicke, 2017; Roediger & Karpicke, 2006a; Rowland, 2014). The beneficial effects of taking tests have been found independent of the provision of subsequent feedback and can be explained by an active recall from memory when trying to produce the correct answer (Roediger & Karpicke, 2006a). Consequently, researchers have advocated the use of tests in educational contexts to improve learning (Dunlosky et al., 2013). Many studies have investigated the testing effect in real-world educational settings such as actual classrooms (for meta-analyses, see Adesope et al., 2017; Schwieren et al., 2017). However, most of these studies did not address a question of highest relevance to practitioners: Do students benefit from answering questions without feedback more that they would from restudying the material? The extant studies on retrieval practice in educational settings either compare taking tests to doing nothing or provide feedback after taking a test, which makes it impossible to differentiate the effects of answering questions from the effects of additional feedback (Moreira et al., 2019). Moreover, a related, equally relevant question is what type of question format should be used to elicit positive effects when practice tests are used without feedback.

Should learners practice test questions or should they restudy learned content in real-world educational settings, whenever feedback on the test is unavailable? Proponents of the testing effect have established the superiority of testing without feedback over restudy in laboratory research (Moreira et al., 2019; Rowland, 2014). Furthermore, findings from simulated-classroom studies suggest that the question format seems to be crucial for testing effects when feedback is withheld: Whereas short-answer testing was found to be superior to restudying, multiple-choice testing was not (e.g., Butler & Roediger, 2007). To our

knowledge, only one study investigated testing effects for different question formats without feedback in real-world educational settings and it has found a similar pattern concerning the question format (Greving & Richter, 2018). These findings are in line with theoretical frameworks of the testing effect that assume retrieval to be the most effective when it is effortful and successful (R. A. Bjork, 1994; Pyc & Rawson, 2009). We argue that the testing effect in educational settings without feedback is underexplored. Moreover, we argue that multiple-choice questions used in simulated and actual classrooms studies did not elicit testing effects because they likely stimulated not enough effortful retrieval to promote retention. The present study examined testing effects in a real-world educational context without feedback and employed short-answer questions as well as multiple-choice questions that were designed to encourage more effortful retrieval. By doing so, it ultimately aims at determining the practical value of testing with different question formats for learning contexts where feedback is not available.

Before explaining the rationale of the study in more detail, we will discuss research on the testing effect from laboratory studies and studies conducted in educational contexts, while also covering the theoretical and practical relevance of investigating the testing effect without feedback. We then outline findings and theoretical frameworks that support the assumption of differential effects of multiple-choice and short-answer questions on the testing effect. In this context, we will also discuss the crucial role of retrievability, that is the classification of items as being mostly successfully retrieved during practice testing or not, for the effect of testing without feedback on retention.

Direct and Mediated Testing Effects

The testing effect describes the beneficial effects of taking tests or quizzes on retention compared to restudying the content that is tested. Retention is usually measured by

a criterial test that takes place sometime after testing and restudying. Roediger and Karpicke (2006a) proposed to distinguish between direct and mediated testing effects. Direct testing effects are thought to be beneficial because of the active retrieval process compared to no retrieval processes when restudying. Mediated testing effects are not the result of retrieval process but instead the result of additional processes that have been triggered by testing. As an example, imagine students attending a lecture and either answering test questions on core content about the lecture or restudying the content. Imagine further that students are informed about their respective practice activity beforehand and that they will also receive feedback in the form of the correct answer. This example illustrates how the test expectancy can lead to heightened attention in the lecture, and feedback can lead to additional study effort, especially for content that has not been retrieved correctly. Expectancy and feedback processes should lead to better retention of the lecture content, mediating the testing effect because they are not attributable to retrieval processes (McDaniel & Little, 2019; Roediger, Putnam, et al., 2011). Conversely, direct testing effects emerge when all mediated testing effects have been controlled, either statistically or by using an appropriate experimental design and setup (Roediger & Karpicke, 2006a). However, a recent review of testing in educational contexts identified that all reviewed studies had employed feedback or had not compared testing to another activity (Moreira et al., 2019). This lack of experimental rigor and potentially mediated testing effects might explain why results obtained in applied research differs from findings in in the laboratory. It is thus important to investigate whether direct testing effects also apply in real-world educational settings.

Aside from theory-driven purposes, investigating the direct testing effect is also relevant for practitioners. Teachers and students should know, for a number of reasons, whether the effort of retrieving the correct answer alone can promote retention. Most

importantly, feedback is time-consuming and the time spent on processing feedback could also be spent on other learning processes. It has been found that feedback can have negative net effects on learning when total learning time was fixed (Hays et al., 2010). Furthermore, feedback also requires learners to invest effort. This constraint seems relevant especially when students use self-testing as a learning strategy during their self-regulated learning activities (Zepeda & Nokes-Malach, 2021). Greving, Lenhard and Richter (2021) found that in such a learning situation, students rarely ever seek feedback, for example by going back to the learning materials, even when they fail to give the correct answer. Practitioners should therefore know whether providing learners with test questions alone is beneficial for fostering retention or if practitioners are required to provide additional feedback in order to render practice tests effective for learning.

Testing Effects in Laboratory and Educational Settings

Distinguishing between direct and mediated testing effects becomes crucial when comparing findings from different experimental settings and for the application of tests in real-world educational contexts. Recent meta-analyses provide evidence for the assumption that testing effects can reliably be found in laboratory studies (Adesope et al., 2017; Phelps, 2012; Rowland, 2014) and in real-world educational contexts (Adesope et al., 2017; Schwieren et al., 2017). The effect sizes determined in these meta-analyses are medium to large for laboratory studies (Cohen's d/Hedges' g ranging from 0.50 to 0.61) and studies in real-world educational contexts (0.54 to 0.67). However, the two experimental settings differ in the number of additional moderators and processes that can be experimentally manipulated and controlled. Laboratory studies usually control for these moderators by means of method designs such as using random assignment to practice conditions, using artificial material that can only be (re)studied in the laboratory, and equating exposure time of practice conditions

(Roediger & Karpicke, 2006a). The extant literature on the testing effect also provides evidence on the impact of specific moderators, for example, the provision of feedback (Rowland, 2014). Consequently, meta-analyses on laboratory studies can provide a good estimate of direct testing effects. In contrast, studies in educational settings often investigate testing effects with the goal of maximizing ecological validity while also using research designs that allow dissociating direct and mediated testing effects or that do not involve strong control conditions. Consequently, past research in these settings often included feedback (e.g., Bell et al., 2015; Carpenter et al., 2009; Cranney et al., 2009; Glass et al., 2008; Goossens et al., 2016; Kromann et al., 2009; Larsen et al., 2009; Leeming, 2002; Leggett et al., 2019; Lyle & Crawford, 2011; Mayer et al., 2009; McDaniel et al., 2007, 2011, 2012, 2013; McDermott et al., 2014; Roediger, Agarwal, et al., 2011; Stenlund et al., 2016, 2017; Trumbo et al., 2016; Weinstein et al., 2016). Furthermore, many studies conducted in real-world educational settings did investigate the effects of direct and indirect testing effects but did not compare it to other activities that are known to promote retention such as restudying (e.g., Batsell et al., 2017; Bell et al., 2015; E. L. Bjork et al., 2014; Downs, 2015; Foss & Pirozzolo, 2017; Glass et al., 2008; Johnson & Kiviniemi, 2009; Khanna, 2015; Lyle & Crawford, 2011; Marsh et al., 2012; Mayer et al., 2009; McDaniel et al., 2007, 2011, 2013; Shapiro & Gordon, 2012; Vojdanoska et al., 2010). To summarize the methodological challenge, research in laboratory settings provides evidence of direct testing effects, whereas a large body of research on testing effects in educational settings investigates testing effects in way that direct and mediated effects cannot be distinguished, or they do not determine direct testing effects using practically relevant control conditions. Consequently, generalizing the beneficial effects of testing alone (without feedback) to real-world educational contexts is hardly possible. This predicament holds especially true for unmediated, direct testing effects.

In an attempt to close this gap, some studies have investigated the direct testing effect in real-world educational settings while minimizing methodological issues. These studies employed a simulated classroom by instructing participants to watch recordings of actual lectures or to read text passages (Butler & Roediger, 2007; Einstein et al., 2012; Kang et al., 2007; Nungester & Duchastel, 1982). Participants then practiced learning content by answering multiple-choice, short-answer questions, or by restudying. Results from these studies indicate that the testing effect depended on the question format. Short-answer testing was beneficial compared to restudying, whereas multiple-choice testing was not (but see Nungester & Duchastel, 1982).

Additional evidence comes from studies that investigate beneficial effects of testing without feedback in comparison to restudying in actual educational settings while using learning material designed for the purpose of the investigation (Bing, 1984; Jaeger et al., 2015; Lipko-Speed et al., 2014). Among these studies, only Jaeger and colleagues reported testing without feedback to outperform a restudy condition. It should be noted that in the presented studies both in simulated and actual classrooms, participants encountered learning material that had no direct practical relevance for them, unlike in real classrooms. It is therefore reasonable to assume that actual learning differs from learning in these studies in terms of motivation, personal involvement, and effort and thus generalizability of these findings to direct testing effects in real-world educational settings is limited.

By far the closest approximation to direct testing effects in educational settings comes from studies comparing the effects of testing without feedback to restudying while using actual course material (Greving & Richter, 2018; Palmer et al., 2019; Thomas et al., 2020). Palmer and colleagues had students that attended lecture sessions either answer open-

ended questions about that lecture or rewatch the lecture. In a criterial test, one week later no differences occurred regarding the retention of lecture content.

Thomas et al. (2020) restructured an existing university course to investigate the beneficial effects of 10-minute practice of multiple-choice questions with and without feedback as compared to 10-minute restudying on the retention of the content of 20-minute lectures. The authors report beneficial effects of testing irrespective of feedback over restudying. However, it is questionable whether this benefit reflects direct testing effects. For one, manipulation of practice after lectures occurred within-subjects after four weeks. Within these four weeks, participants repeatedly experienced the same condition twice per class and each class twice per week, resulting in approximately 16 practice sessions of the same kind before the conditions were altered. Thus, students were able to anticipate the type of practice for which an extra effort would be helpful. This is especially worrying since practice test performance counted toward students' final grades and students might have altered their learning behavior and strategies, because they were motivated to get good grades. In sum, the results found by Thomas et al. might well be the result of indirect testing effects.

Greving and Richter (2018) studied the testing effect in the context of an existing university lecture. After attending lecture sessions, students were assigned randomly to review conditions. Practice was limited to the last 10 min of a 90-minutes lecture without the possibility to revisit the material or receive feedback. Performance in practice sessions and on surprise criterial tests did not count toward students' final grades. Practice questions in the lab study by Butler and Roediger (2007) were selected because of their high retrievability (i.e., items that were answered correctly most of the time and are thus easy in terms of item difficulty), whereas Greving and Richter (2018) assessed the retrievability in the practice sessions. More precisely, they used the difficulty of the items used in the practice sessions to

define groups of items that differed in retrievability. Findings across different retention intervals were in line with the results from most simulated-classroom studies. Practicing short-answer questions led to more retention than restudying, whereas practicing multiple-choice questions did not. However, a testing effect for short-answer questions occurred for highly retrievable items only.

The findings from most simulated-classroom studies and Greving and Richter (2018) that multiple-choice testing is not better than restudying seems to contradict the findings from meta-analyses showing robust testing effects for multiple-choice questions (Adesope et al., 2017; Rowland, 2014). However, these meta-analyses notably used laboratory studies only (Rowland, 2014) or to a large extent (88%; Adesope et al., 2017) and comparisons between laboratory studies and studies in educational contexts might be difficult for the reasons outlined above. Furthermore, the results of most simulated-classroom studies and the study by Greving and Richter also seem to be at odds with some studies conducted in real-world educational settings, suggesting that multiple-choice questions elicit a testing effect (Thomas et al., 2020). In their review of the literature on testing effects in educational contexts, Moreira et al. (2019) provided evidence for the existence of testing effects for multiple-choice questions and even an advantage of multiple-choice over short-answer testing. However, Moreira et al. also acknowledged that all reviewed studies provided feedback or used a weak control condition (e.g., no learning activity). Whereas provision of feedback makes it impossible to differentiate between direct or mediated testing effects, the activity in the control condition determines whether found differences between testing and control conditions bear practical implications.

To summarize, studies investigating direct testing effects in educational settings – in a way that such effects can be isolated from mediated testing effects – are scarce and often

found short-answer testing not to be superior to restudying (Bing, 1984; Lipko-Speed et al., 2014; Palmer et al., 2019). Most simulated-classroom studies as well as an investigation in an existing university course by Greving and Richter (2018) support the idea that in educational settings, answering short-answer questions compared to restudying seems to be beneficial for learning, which is similar to findings from laboratory research. Answering multiple-choice questions, however, has only been shown to be superior in laboratory research or when other mediating factors such as feedback were present. Studies from applied educational contexts when controlling for these factors found no difference between multiple-choice testing and restudying, except when practice tests were likely to elicit indirect testing effects (Thomas et al., 2020).

Theoretical Accounts on the Testing Effect and the Role of Effortful and Successful Retrieval

Reviewing theoretical frameworks of the testing effect, such as the desirable difficulty framework (R. A. Bjork, 1994), the new theory of disuse (R. A. Bjork & Bjork, 2006), and the retrieval effort hypothesis (Pyc & Rawson, 2009), can help to understand why the testing effect might differ between question formats and also highlights the relevance of another potential moderator, retrievability of the knowledge that is tested. The available theories all incorporate the assumption that the retrieval process causes the direct testing effect and that more effortful retrieval leads to better retention of the learned material, which is why testing should lead to better retention than restudying. Furthermore, retrievability plays a crucial role in all of these theoretical frameworks. When the correct information cannot be retrieved from memory, testing has no benefits compared to restudying (e.g., Jang et al., 2014). Retrievability matters most when practice testing is used without feedback, because in this situation, only learners who retrieve the correct answer are likely to strengthen the memory trace of the to-be-learnt information, enhancing its later retrievability. In

contrast, learners who are unable to retrieve the correct answer essentially learn nothing because they can benefit neither from the direct testing effect nor from using feedback to encode (or be reminded of) the correct answer.

Multiple-choice and short-answer questions differ in the processing that results in correct answers and—given these theoretical underpinnings—should consequently lead to different testing effects (e.g., Karpicke, 2017). The beneficial effects of testing seem to depend largely on the depth of processing involved in the retrieval process (Carpenter & Delosh, 2006; Rowland, 2014). This fundamental account of the testing effect also provides an explanation for different test formats being more or less beneficial. Free recall and short-answer questions often (although not always) require deeper processing than multiple-choice questions. Depending on the knowledge of the test-taker, multiple choice questions may be solved with different strategies, including the elimination of distracters, which involves knowledge retrieval (Little, Frickey, & Fung, 2019). However, if the correct response directly matches learnt information, multiple choice questions can be answered based on the recognition of the correct response only and often do not active retrieval. In that case, multiple choice questions are less likely to lead to a direct testing effect than short answer tests. Glover (1989) has shown that free and cued recall lead to more pronounced testing effects than recognition tasks. One additional difference between the question formats is that researchers and practitioners do not always use verbatim repetitions of retrieval practice in criterial tests and exams. Instead, questions are used that ask for related information. Previous studies suggest that these questions may lead to impaired retrieval, a phenomenon dubbed retrieval-induced forgetting and is most pronounced when practicing multiple-choice questions (Carroll et al., 2007). However, research suggests that context variables exist that prevent students from experiencing retrieval-induced forgetting such as delay between initial

test and criterial test (Chan, 2009) and learning coherent information (Little et al., 2011). Furthermore, prior investigations of multiple-choice questions in real-world educational settings yielded findings that are inconsistent with retrieval-induced forgetting induced by multiple-choice questions (E. L. Bjork et al., 2014).

Based on the assumption that the adverse effects occur because of shallow processing and mere recognition tasks, deep processing and effortful retrieval should provide beneficial testing effects. Constructing test items in a manner that triggers active retrieval processes may lead to higher direct testing effects (Little et al., 2012). The lack of direct testing effects for multiple-choice testing could therefore be explained by the usage of multiple-choice questions that do not invoke active retrieval.

Rationale of the Present Study

Based on our discussion of previous research on the testing effect, investigations in educational settings should target the direct testing effect, that is, the beneficial effects of testing compared to restudying that are mainly attributable to practicing learned content, similar to laboratory experiments investigating the testing effect in laboratory research (e.g., Roediger & Karpicke, 2006b). The aim of the present study was to re-examine differential direct testing effects in a university course and to conceptually replicate the field experiment by Greving and Richter (2018). To ensure a high level of internal and external validity, we took several measures in the original study and its conceptual replication. First, participants who had just attended a lecture were assigned randomly to either answering short-answer questions, multiple-choice questions, or reading summarizing statements about core lecture content. No feedback was provided in either condition. This random assignment following each lecture was done because previous research informed participants beforehand whether the learning content would be tested and this information might have an effect on initial study

behavior (e.g., Batsell et al., 2017; Daniel & Broida, 2004; Trumbo et al., 2016). Other studies simply varied conditions between classes, rendering it impossible to differentiate effects of the manipulation from effects of the classes (Batsell et al., 2017; Cranney et al., 2009; Khanna, 2015; Leeming, 2002; Mayer et al., 2009; Vojdanoska et al., 2010). Second, the present study aimed at equating exposure time of all review conditions. Previous research allowed for multiple testing, resulting in differing exposure times and making it impossible to estimate the value of each testing occasion (Downs, 2015; Johnson & Kiviniemi, 2009; McDaniel et al., 2012; Stenlund et al., 2016; Trumbo et al., 2016; Wiklund-Hörnqvist et al., 2014; Yong & Lim, 2016). Third, in an attempt to approximate learning assessments in a realistic setting, the criterial test contained not only questions identical to those used on the practice tests but also alternate (near-transfer) versions that asked for the same knowledge in a different way.

In the original study by Greving & Richter (2018), retention for learning content was assessed using three unannounced criterial tests assessing practiced and unpracticed content. The procedure in the present study deviated from the original study by administering only one criterial test, given that Greving and Richter found the same pattern of results across all criterial tests. The procedures also differed in retrieval difficulty in answering multiple-choice questions. Multiple-choice testing has been found to not produce testing effects because of the comparably shallow processing and effortless retrieval needed to answer these questions. To investigate the potential of practicing multiple-choice questions, care was taken to foster productive retrieval in multiple-choice questions. To this end, multiple-choice questions were modified in a way that required respondents to mark every option which correctly answered a question (multiple-response questions; for other terms, see Verbic, 2012). Instead of only recognizing the correct answer among the options, each or none of the provided response

options might be correct and respondents needed to retrieve all information to individually decide for each option whether it is correct or not. Multiple-response questions have been found to increase item difficulty by reducing guessing effects (Kubinger & Gottschall, 2007). Given that all response options can be correct or incorrect at the same time, response options are more competitive than simple multiple-choice questions. It has been shown that competitive responses options trigger more productive retrieval processes than their non-competitive counterparts (Little et al. 2012). The present study investigate whether these possible effects also play out in a realistic learning environment.

We expected to conceptually replicate the findings of Greving and Richter (2018). In particular, we expected a positive effect of short-answer testing and no positive testing effect for multiple-choice questions, and we expected retrievability to similarly influence the testing effect such that higher levels of retrievability should yield bigger testing effects.

Method

Participants

Participants were 53 undergraduate students in their second semester, most of them female (74%) and psychology majors (98%). They participated in at least one lecture session and in the criterial test. All students gave their informed and written consent prior to participation. Participants' age ranged between 19 and 47 years with a mean age of 23.51 (SD = 6.53) years. To be admitted to a psychology program at a German university, students must have achieved very good grades in the university entrance exam (the German Abitur, i.e. the school leaving certificate of the academic-track high school). Therefore, it may be assumed that the overall achievement level in the sample is high. The course was an introductory psychology lecture aimed at students in their first terms, teaching basic principles of cognitive psychology. The university is a regular public German university (Universität) and

does not differ much from other public regular universities in terms of subject range, degrees offered, and teaching requirements. For the reported study, no ethics approval was required per the guidelines of the university or national guidelines. However, the study has been conducted in an ethical and responsible manner and is in full compliance with all relevant codes of experimentation and legislation.

Materials

Test Questions and Restudy Statements

The content of five lecture sessions of an introductory lecture on the psychology of learning, emotion, and motivation was surveyed and 24 information units per session were identified. For each information unit, one summarizing statement, one short-answer question, and one multiple-choice question were created. Summarizing statements were created by summarizing key concepts of the information unit in one sentence (e.g., "Extending emotional reactions that have been learned in response to certain stimuli to new stimuli that are similar to the original stimuli is called stimulus generalization"). Short-answer questions were created by asking for explications of the key concepts in the information unit (e.g. "Please describe the key concept of stimulus generalization?"). Multiple-choice questions were created by reformulating the short-answer question and adding four response options plus a "None of the above" option (e.g., "What describes the key concept of stimulus generalization? A: Stimulus generalization occurs when emotional responses extend to stimuli that are similar to conditioned stimuli, B: Stimulus generalization occurs in the context of operant conditioning, C: The opposite of stimulus generalization is aversive counter-conditioning, D: Stimulus generalization occurs with new stimuli."). For multiple-choice questions, any number from 0 to 4 response options could be correct. The "None-of-the-above"-option was never the correct answer as research has shown that using multiple-

choice questions where “None of the above” is the correct answer can lower the effectiveness of practice test (Blendermannet al., 2020; Odegard & Koen, 2007).

Review Materials

For each of the five lecture sessions, eight information units were randomly drawn from the 24 information units prepared for this session. Based on the selected information units, review materials were prepared for each lecture session. The review materials consisted of a one-page questionnaire asking for basic demographic information and two pages of review items corresponding to the selected information units, consisting of either (a) eight summarizing statements (restudy condition), (b) eight short-answer questions (testing, short-answer questions), or (c) eight multiple-choice questions (testing, multiple-choice condition). In all three versions, information units were presented in the same order with four information units on each page.

Criterial Test

The criterial test was constructed of questions based on the pool of 24 information units determined for each of the five lecture sessions. For providing some near transfer in the assessment of learning, and therefore making the criterial tests more similar to realistic learning assessments, the pool of questions was expanded by creating alternate versions of the questions used in the review material. Alternate questions were created by asking for the key concepts in a different way (e.g., “What term describes the extension of reactions that have been learned in response to certain stimuli to new stimuli that are similar to the original stimuli?”). For each information unit, an alternate short-answer question and an alternate multiple-choice question were created.

The criterial test consisted of three components: (a) questions corresponding to information units included in the review materials, (b) questions corresponding to

information units not included in the review materials, and (c) alternate questions that corresponded to information units but were not identical to questions included in the review materials. Furthermore, the test consisted of short-answer questions and multiple-choice questions in equal proportions. Two versions were created (Versions A and B) by altering the order of questions and the question format (i.e., multiple-choice questions vs. short-answer questions) of the same question between criterial test versions so that all multiple-choice questions in Version A were short-answer questions in Version B and vice versa.

Scoring

Short-answer questions were scored with 1 (correct response) vs. 0 (incorrect or missing response). Two independent raters scored all responses to short-answer questions. Inter-rater reliability was high across all lectures and criterial tests (576 observations, Cohen's $\kappa = .80$). Thus, scores from only one rater were included in the analyses. Multiple-choice questions were scored using the PS_{50} procedure. The PS_{50} is a partial credit scoring algorithm that awards 0.5 points whenever more than half of the ticks are set correctly and awards 1 point when all ticks are set correctly (Lahner et al., 2018). As an example, consider a question with response options "A"–"D" and "A" and "B" being correct. Participant 1 ticks "A", resulting in 75% of the ticks set correctly, Participant 2 , ticks "A" and "C", resulting in 50% of the ticks set correctly, and Participant 3 ticks "A","B", and "C", resulting in 25% of the ticks set correctly. Application of the PS_{50} procedure results in Participant 1 receiving 0.5 points whereas Participants 2 and 3 both would receive 0 points. We used PS_{50} scores to calculate item properties that were used as predictors in modeling criterial test scores. However, for the analysis of criterial test scores, PS_{50} scores were furthermore dichotomized to be compatible with short-answer scores. In the example, Participant 1 would thus receive a

dichotomized score of 1 and Participants 2 and 3 a dichotomized score of 0. The performance scores based on both question types in the criterial test served as the dependent variable.

Procedure

The study was conducted in a weekly introductory psychology lecture on the basic principles of learning, emotion, and motivation. Lecture Sessions 4, 5, 7, 8, & 9 served as focal sessions. Manipulation of the review condition (testing with multiple-choice or short-answer questions) took place in these five lecture sessions. The criterial test, which assessed content learning taught in the sessions, was administered unannounced in Session 10.

In each focal session, the last 10 min of the lecture session was reserved for manipulation of the review condition. Students were told that the study offered the opportunity to review the lecture content and that participation might be helpful for their own learning activities. Nevertheless, participation was voluntary. Students were allowed to leave the lecture hall after the end of the regular lecture. Research assistants then administered the review materials, assigning participants randomly to one of the three review conditions (testing with multiple-choice questions, testing with short-answer questions, or restudying). Participants first filled in basic demographic information. They were then given 4 min to complete each page of the two pages of items for review. Finally, participants were thanked for their participation, and the materials were collected.

All students who were present in Session 10 were allowed to take the criterial test, irrespective of previous participation in the study. The two versions (Version A and B) of the criterial test were administered in an alternating way so that participants sitting next to each other received different versions. Students were allowed 45 min to complete the test and could leave when they finished.

Design

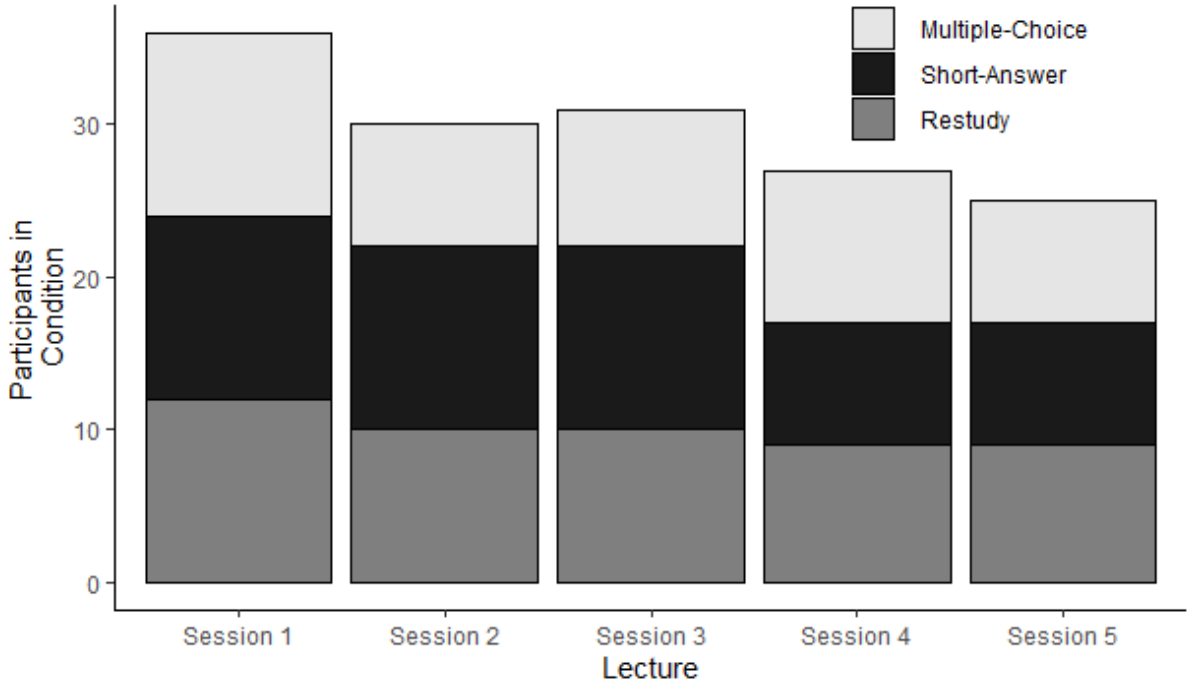
The design was a one-factorial within-subjects design with the independent variable review condition (multiple-choice test, short-answer test, restudying). Moreover, each participant received one of two versions of the criterial test, which differed in format (short-answer vs. multiple-choice question) and the order of questions. The dependent variable was the performance (probability of a correct response) on the multiple-choice and short-answer questions in the criterial tests.

The design was implemented by randomly assigning participants in each of the focal lecture sessions to one of the three review conditions. Likewise, participants were assigned to one of the two test versions of the criterial tests administered. Figure III.1 depicts the number of participants that were assigned to each review condition in the five lecture sessions. The random allocation led to equal distributions of participants across review conditions. Similarly, participants were almost evenly distributed to the criterial test versions (Version A: $n = 25$; Version B: $n = 28$). Of the 53 participants present in the criterial test, 14 had received a review condition in all five lecture sessions, eight participants only in four lecture sessions, five participants only in three lecture sessions, 13 participants only in two lecture sessions, and six participants only in one lecture session. Seven participants present in the criterial test did not experience any review condition following a lecture session. Given these distributions of participation and the random allocation of review conditions, participants on average answered equal amounts of questions in the criterial test that have been practiced using short-answer, questions, multiple-choice questions, and restudy (short-answer questions: $Mdn = 6$, $M = 5.74$, $SE_M = 5.97$; multiple-choice questions: $Mdn = 6$, $M = 5.19$, $SE_M = 5.36$; restudy: $Mdn = 6$, $M = 5.74$, $SE_M = 5.97$). We assumed that missing data were missing completely at

random and thus inferences could proceed by analyzing only the observed data (Ibrahim & Molenberghs, 2009).

Figure III.1

Participation by Condition and Session



Availability of Data and Materials

Materials in the form of items used in the practice sessions and criterial tests as well as data and analysis scripts underlying the results reported in this study are provided in the repository of the Open Science Framework (https://osf.io/xgc4e/).

Results

We used generalized linear mixed-effect models (GLMMs) with a logit-link function (Dixon, 2008) to analyze correctness of provided answers in the criterial test. Considering the binary nature of this outcome variable, generalized mixed-effect models have been shown to be superior to alternative approaches such as ANOVAs of aggregated scores

(Dixon, 2008; Jaeger, 2008). Furthermore, mixed-effect models have many advantages compared to other analysis techniques (e.g., see Baayen et al., 2008; Richter, 2006), including an economic approach to handling missing data.

We adopted a fully Bayesian approach for analyzing our data for two reasons. The first reason is the possibility to incorporate prior knowledge about parameters into model estimation. The present study is a conceptual replication of a previous study (Greving & Richter, 2018). We therefore used parameter estimations from the earlier study as priors for the estimation of Bayesian mixed-effect models.

The second reason to use Bayesian statistics is that it enabled us to obtain probability estimates for parameter values and to estimate the probability that a given parameter has no influence on the outcome variable. Therefore, the Bayesian approach allows to draw conclusions in the context of a null effect.

We used the brms package (Bürkner, 2017) to estimate Bayesian multilevel models. We report model parameters and the corresponding 95% credible intervals (CIs). Depending on the direction of the hypothesis, we estimated one of three Bayes factors using functions from the bayestestR package (Makowski et al., 2019) to test hypotheses. The first Bayes Factor compares the fit of a model that assumes a positive effect of a predictor on the dependent variable to a model that assumes no effect or a negative effect of a predictor on the dependent variable (BF_{pos}; values larger than 1 indicate evidence in favor of positive effects). A predictor needed to exceed $\beta = 0.2$ to be positive, which corresponds to a 5% increase in the probability of answering criterial test questions correctly ($OR = 1.22$). The second Bayes factor examines a negative effect of a predictor ($\beta = -0.2$; BF_{neg}; values larger than one indicate evidence in favor of negative effects) and is consequently associated with a 5% decrease in the probability of answering criterial test questions correctly ($OR = 0.82$). The

third Bayes factor compares the fit of a model that assumes the predictor to have no effect on the dependent variable (null effect) to a model that assumes the predictor to be non-zero (BF_{01}; values larger than one indicate evidence in favor of the null hypothesis).

Separate models were estimated to examine the testing effect based on short-answer questions and the testing effect based on multiple-choice questions. In each model, the testing condition was compared to the restudy condition that involved reading the summarizing statements that provided the correct answer (dummy coded: testing = 1, restudy = 0). We additionally tested whether the testing effect depended on the retrievability of learned information by including two dummy-coded predictors that contrasted items of medium retrievability and low retrievability with items of high retrievability as the reference condition. We examined whether higher retrievability rates were associated with a larger testing effect. To construct this predictor, we grouped the short-answer questions and the multiple-choice questions (PS_{50} scores) separately into three equally-sized, ordered categories (tertiles) according to their difficulty in the practice tests. To avoid distortions from extreme values, we discarded the lowest and the highest 5% of the distribution before the grouping. Data points (i.e., participant-item combinations) are roughly equally distributed across the three difficulty levels (short-answer: low = 108, medium = 86, high = 98, multiple-choice: low = 89, medium = 89, high = 82). Item difficulties from multiple-choice questions were corrected for guessing.

For each of the two item types (short-answer and multiple-choice questions), grouping resulted in three categories of items with high (short-answer questions: item difficulties from 54% to 78%; multiple-choice questions: 48% to 80%), medium (short answer questions: 30% to 53%; multiple-choice questions: 33% to 47%), and low retrievability (short answer questions: 4% to 29%; multiple-choice questions: 8% to 32%).

Finally, the models included the interaction of retrievability with testing vs. restudying. All predictors and their interactions were entered simultaneously in the models. Participants and test items were included as random effects (random intercepts) in all models which means that differences between participants as well as differences between items (that may arise due to items stemming from different lectures or from the ordering of items) are considered as having an effect on retention.

Priors were obtained from Greving and Richter (2018) who investigated a similar research question and used an almost identical method (although in a different lecture on a different topic and with different practice and criterial tests). The authors reported individual models for the testing effect based on short-answer questions and also for the testing effect based on multiple-choice questions. We extracted parameter estimates from both models and used them as priors for fixed effects and intercepts in the Bayesian models that estimated the testing effect for short-answer questions and for multiple-choice questions. We ran 100,000 iterations per model and set the thinning rate to 100.

Table III.1

Model Parameters

	Short-Answer Questions			Multiple-Choice Questions		
Parameter	β	Lower 95% CI boundary	Upper 95% CI boundary	β	Lower 95% CI boundary	Upper 95% CI boundary
Intercept	0.28	-0.44	1.06	0.22	-0.49	0.94
Testing	0.57	-0.21	1.41	-0.54	-1.33	0.18
Low retrievability	-1.28	-2.42	-0.23	-0.80	-1.81	0.26
Medium retrievability	-0.19	-1.12	0.70	-0.48	-1.56	0.56
Testing x Low retrievability	0.57	-0.62	1.74	0.98	-0.16	2.07
Testing x Medium retrievability	-0.11	-1.37	1.18	0.77	-0.64	2.28
$n_{\text{Participants}}$		43			43	
n_{Items}		29			29	

Note. Model parameters for the models estimating the effect of testing with short-answer questions and multiple-choice questions and retrievability on learning performance in the criterial test

Effects of Testing with Short-Answer Questions

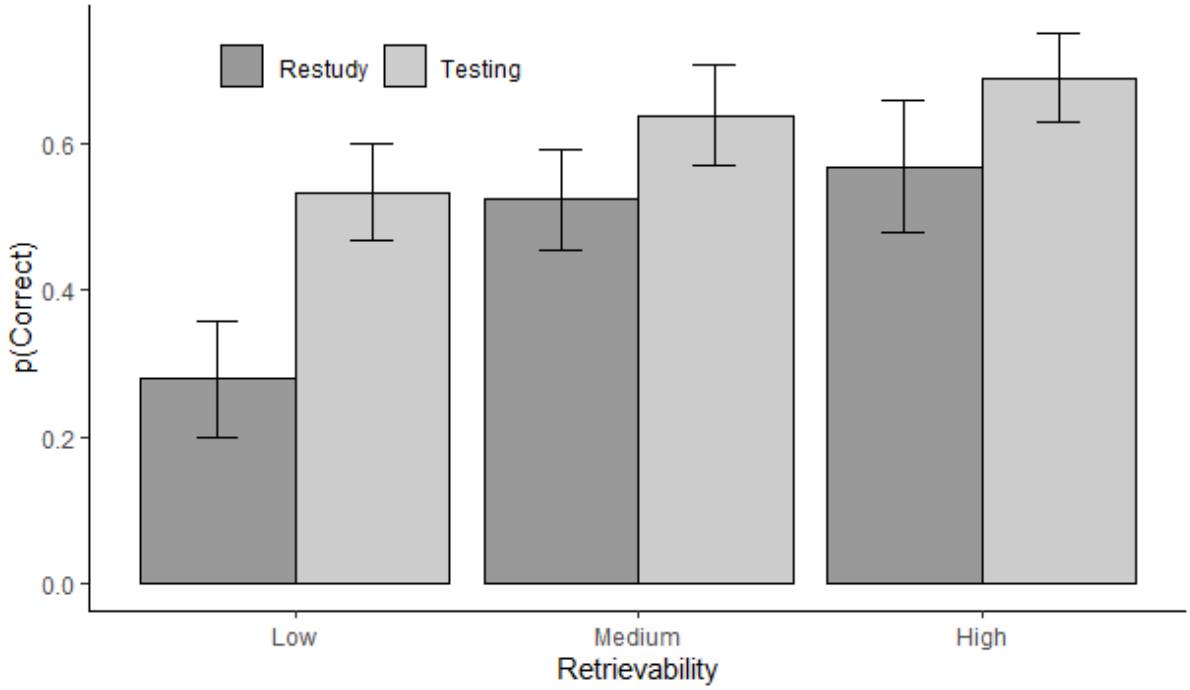
The model estimates for the effects of testing with short-answer questions can be found in Table III.1 (left columns). This model revealed moderate evidence for a positive effect of testing (β = 0.57, 95% CI [-0.21, 1.41], BF_{pos} = 4.01). Contrary to our predictions, we found moderate evidence for a positive interaction of testing vs. restudying with the predictor comparing low to high retrievability (β = 0.57, 95% CI [-0.62, 1.74], BF_{neg} = 0.29, BF_{pos} = 3.43). However, we found evidence for a negative interaction with the predictor comparing medium to high retrievability but also stronger evidence for the absence of any

interactions of testing with this predictor (β = -0.11, 95% CI [-1.37, 1.18], BF_{neg} = 2.14, BF_{01} = 11.30). The predicted values resulting from this model are shown in Figure III.2.

Most importantly, planned contrasts revealed a general testing effect across all retrievability levels that increased the probability of answering correctly in the criterial test by 17% (95% CI [5.63, 27.00], *OR* = 2.06). The testing effect was most pronounced for items with low retrievability (+27%, 95% CI [7.27, 43.66], *OR* = 3.12). For medium (+11%, 95% CI [-10.15, 32.20], *OR* = 1.59) and high retrievability (+13%, 95% CI [-4.81, 31.89], *OR* = 1.78) the same trend emerged, but 95% CI included zero, providing no clear evidence for a benefit of testing compared to restudying.

Figure III.2

Effects of Short-Answer Questions as Probability of Correct Responses in Criterial Test Items



Note. Testing with short-answer questions. Probability of correct responses in criterial test items (back-transformed from the logits in the GLMM) by retrievability and learning condition (testing vs. restudy). The median of posterior distribution and 95% credible interval are presented.

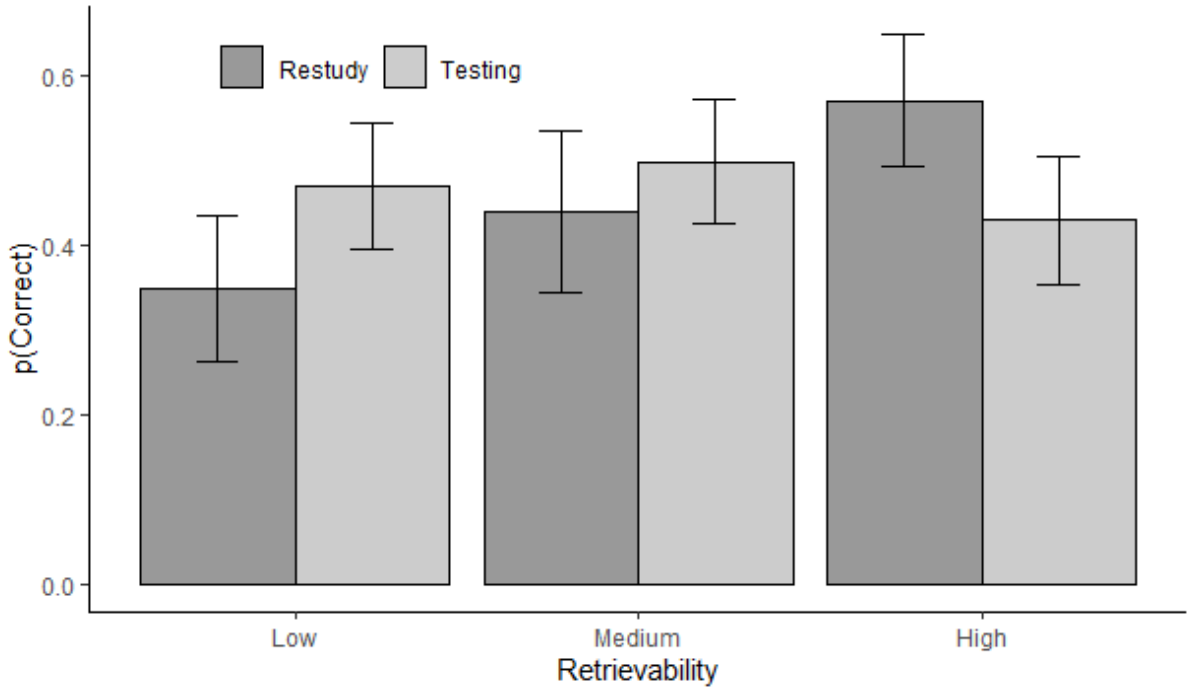
Effects of Testing with Multiple-Choice Questions

The model estimates for the effects of testing with multiple-choice questions can be found in Table III.1 (right columns). This model revealed moderate evidence for a null effect of testing (β = -0.54, 95% CI [-1.33, 0.18], BF_{01} = 5.06). Furthermore, we found moderate evidence against an interaction of testing vs. restudying with the predictor comparing low to high retrievability (β = 0.98, 95% CI [-0.16, 2.07], BF_{01} = 2.15) and strong evidence against an interaction of testing vs. restudying with the predictor comparing medium to high retrievability (β = 0.77, 95% CI [-0.64, 2.28], BF_{01} = 6.40). The predicted values resulting from this model are shown in Figure III.3. Planned contrasts revealed that the probability of

answering correctly in the criterial test was not increased with multiple-choice questions. The increase was 1% and the confidence interval lay almost symmetrical around zero, providing no evidence for an effect of testing (95% CI [-9.48, 11.48], *OR* = 1.04). Furthermore, this lack of evidence for any difference between testing and restudying was independent of retrievability given that the 95% CI included zero on all retrievability levels (high retrievability: -12%, 95% CI [-30.99, 3.64,], *OR* = 0.59; medium retrievability: +5%, 95% CI [-20.83, 28.23], *OR* = 1.25; low retrievability: +11%, 95% CI [-10.99, 29.58], *OR* = 1.57).

Figure III.3

Effects of Multiple-Choice Questions as Probability of Correct Responses in Criterial Test Items



Note. Testing with multiple-choice questions. Probability of correct responses in criterial test items (back-transformed from the logits in the GLMM) by retrievability and learning condition (testing vs. restudy). The median of posterior distribution and 95% credible interval are presented.

Discussion

The present study re-examined differential testing effects in an existing educational context by comparing retrieval practice using effortful multiple-choice questions and short-answer questions to restudying. We adapted the paradigm by Greving and Richter (2018) that aimed at minimizing methodological problems while investigating the direct effects of testing that are mainly attributable to practicing learned content—similar to the testing effect investigated in laboratory research (e.g., Roediger & Karpicke, 2006b). In doing so, we tested the claim that answering practice questions—even without feedback—benefits retention more than restudying in a real-world educational setting.

One main finding was a testing effect for practice tests based on short-answer questions. Subgroup analyses similar to the previous study by Greving and Richter (2018) revealed that this effect was only detectable in the subgroup of learning content that was most difficult to retrieve from memory. Another main finding pertains to practice tests based on multiple-choice questions. We found evidence for the absence of any testing effect for multiple-choice questions.

Our findings on the beneficial effects of answering short-answer questions are in line with laboratory research that investigated direct testing effects in educational contexts by means of a simulated classroom (Butler & Roediger, 2007; Einstein et al., 2012; Nungester & Duchastel, 1982). The findings differ, however, from the study by Greving and Richter (2018) that examined these effects in an existing university course. In the present study, as well in most simulated-classroom studies, short-answer testing was superior to restudying. A beneficial effect of short-answer testing only emerged with low-retrievable items, which stands in contrast to our hypotheses and the findings by Greving and Richter (2018). The latter only found testing effects for highly retrievable items, that is, items that were answered mostly correct in the testing subsequent to learning. These discrepancies of results can be explained by differing retrievability levels between the studies. Butler and Roediger, for example, explicitly used only items that were found to be highly retrievable (M = 68%). This limitation of items can be justified by theoretical accounts on the testing effect, mainly the bifurcation model, which states that the superiority of testing compared to restudying depends on the amount of successfully retrieved items in the testing condition (Halamish & Bjork, 2011; Kornell et al., 2011; Rowland, 2014). The findings from Greving and Richter (2018) are consistent with these accounts when considering that testing effects only emerged for items that had the highest retrievability rates (46%–81%). Notably, retrievability in applied

educational contexts depends on factors that are difficult or impossible to experimentally control (e.g., the way it was originally taught, the complexity of the topic, other overlapping/related information that could result in a misconception). Presumably, this is why retrievability rates in the present study (M = 41%, SD = 23%) were higher than in the study by Greving and Richter (M = 37%, SD = 20%; $t(366.87) = 3.09$, $p = .002$). These higher retrievability rates might explain why a general testing effect appeared in the present study and not in the study by Greving and Richter.

Higher overall retrievability rates might also explain the finding that the testing effect was only detectable in the subgroup with lowest retrievability rates. This claim is backed by the retrieval effort hypothesis (R. A. Bjork, 1994; Pyc & Rawson, 2009), which assumes testing to be the most beneficial approach when retrievability is most difficult but successful. Given that higher retrievability rates are synonymous with more successful testing, it is reasonable to assume that the subgroup comprised of the hardest items yields the strongest testing effects.

Our findings showing the lack of beneficial effects of answering multiple-choice questions are in line with studies that investigated the direct testing effects in simulated and actual classrooms (Butler & Roediger, 2007; Einstein et al., 2012; Greving & Richter, 2018; Kang et al., 2007, Experiment 1). Furthermore, the present study provides evidence for the absence of direct testing effects for multiple-choice questions. In conjunction with prior research, these findings support the assumption that practicing retrieval by only answering multiple-choice promotes retention of learning content no more than restudying. However, it should be noted that there are studies from both simulated (Nungester & Duchastel, 1982) and actual classrooms (Thomas et al., 2020) that suggest a different conclusion. Although methodology in these studies varies from the present study, further research is necessary to

determine under what conditions multiple-choice testing can elicit direct testing effects in applied educational settings. As mentioned in the introduction, separate mechanisms might explain the present findings as well as inconsistencies with previous research.

One potential explanation is provided by the assumption that multiple-choice questions fail to promote retrieval that requires much cognitive effort because they can (at least sometimes) be answered correctly by mere recognition or by educated guessing. Consistent with current theories on the testing effect, less effort in retrieving is associated with less retention of the retrieved content (R. A. Bjork, 1994; R. A. Bjork & Bjork, 2006; Pyc & Rawson, 2009). Notably, these results were obtained despite our attempt at inducing more productive retrieval with multiple-choice questions, by allowing multiple response options to be correct (rather than having just one correct response options plus distracters).

With the multiple-response questions, we introduced a question format that offered more productive retrieval opportunities and higher retrieval effort demands than traditional multiple-choice questions. However, it is also possible and compatible with our results that retrievability is not the main attribute that determines whether multiple-choice or multiple-response questions are beneficial or not. Smith and Karpicke (2014) compared multiple-choice to short-answer and a hybrid question format in which participants had to attempt to recall the correct answer to a question (similar to short-answer questions) before they were able to select a response option. Over the course of three experiments, no differences occurred between question formats but when retrievability of the tested items was increased, the short-answer and hybrid question format outperformed the multiple-choice format. Thus, it might be that retrievability had differential effects on these question formats, for example because test takers rely more on recognition than on retrieval when responding to multiple-choice questions.

The second explanation pertains to the nature of multiple-choice questions by which respondents are required to process response options and therefore are presented with erroneous information. Roediger and Marsh (2005) showed that multiple-choice testing may lead participants to answer later criterial tests with false information, but feedback following multiple-choice questions might reduce these negative effects and increase beneficial effects (Butler & Roediger, 2008).

To summarize, answering multiple-choice questions might not have led to effortful retrieval of information from memory or even caused test takers to remember erroneous information. In principle, both processes might explain why answering multiple-choice questions about a 90-minute lecture was equally effective as restudying key concepts of that lecture. Apart from other methodological issues, previous research that demonstrated multiple-choice testing to be superior to restudying even without feedback, presented learning material for a mere 15 (Nungester & Duchastel, 1982) or 20 minutes (Thomas et al., 2020) before restudy or tests took place. It is possible that with more to-be learned information and longer intervals between a first presentation of the information and subsequent practice, restudying might grow in value to learners because it can help learner summarize and organize information after this time. Conversely, with less information and less time between the first presentation and revisiting the learning content, additional restudy might not seem valuable to learners, which might also reduce learners effort invested in restudying. Additional research should thus investigate whether the amount of information and time spent on the first study of this information moderates direct testing effects for multiple-choice questions in educational contexts.

Additionally, based on a review of research conducted on testing effects in educational context, Butler (2018) concluded that multiple-choice question produce reliable

testing effects whenever they are simple in terms of possible response options and allowing only one correct response option. The operationalization of multiple-choice questions in the presented study was founded on theoretical principles but disregarded these considerations which might be the cause for not finding a testing effect for multiple-choice questions. It should however be noted that Greving and Richter (2018) used a simpler multiple-choice format in a similar paradigm as in the present investigation and did also not find any evidence for direct testing effects.

The present study aimed at contributing to understanding the gap between testing effects in controlled laboratory settings and testing effects found in educational settings. The experimental design in a field study is a strength of the present study, but the method also presents some limitations. A limitation that this study shares with other studies investigating learning in natural settings are external influences. Compared to laboratory experiments, external influences potentially play a much greater role in a field settings. The extent that other factors in the present study (e.g., metamemorial, metacognitive, or motivational factors) influenced learning behavior during lectures and review conditions, when taking the criterial tests, or in the days and weeks between the lectures and the criterial tests is unknown. For example, we cannot rule out that some students looked up the answers to the questions on their own (after the practice session), although it seems unlikely that many students did that, given that they did not take the questions home and given that Greving et al. (2021) found that feedback-seeking rarely occurred in a similar learning situation. Participation in the study in each of the lectures and the ensuing practice sessions was voluntary, which might have caused selection effects. However, these selection effects likely affected all experimental conditions to the same extent because participants were unaware of the review condition. Finally, for practical reasons, we were unable to disentangle effects of different types of

questions used in the practice tests on different types of questions in the criterial test; we simply could not include a sufficient number of questions of each type to achieve a reliable and representative assessment of the to-be-learnt knowledge from each lecture session. Likewise, it is an open question whether and how testing with different question types can promote transfer and application questions of a more complex nature, which are also common in typical assessments at the university. We would expect that again, only questions that require a certain degree of cognitive effort would be suitable to improve performance in such questions, compared to restudy.

Prior research has examined the testing effect in educational settings that were inferior to laboratory research in terms of the extent of experimental control. In contrast, the present study provides clear evidence for the claim that answering short-answer questions only once and without feedback benefits retention of learning compared to restudying key points of the lecture. This study also provided evidence that presenting students with multiple-choice questions is equally effective as restudying key points of the lecture. Given these findings, we advise practitioners to use short-answer testing rather than multiple-choice testing to foster learning in university teaching and that practitioners encourage learners to use short-answer questions when learning on their own.

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

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

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

Bing, S. B. (1984). Effects of testing versus review on rote and conceptual learning from prose. *Instructional Science, 13*(2), 193–198. https://doi.org/10.1007/BF00052385

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. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). MIT Press.

Bjork, R. A., & Bjork, E. L. (2006). Optimizing treatment and instruction: Implication of a new theory of disuse. In L-G. Nilsson & N. Ohta, *Memory and society: Psychological perspectives*. Psychology Press.

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10/gddxwp

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

Butler, A. C., & Roediger, H. L. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition*, *36*(3), 604–616. https://doi.org/10.3758/MC.36.3.604

Carpenter, S. K., & Delosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory & Cognition*, *34*(2), 268–276. https://doi.org/10.3758/BF03193405

Carpenter, S. K., Pashler, H., & Cepeda, N. J. (2009). Using tests to enhance 8th grade students' retention of U.S. history facts. *Applied Cognitive Psychology*, *23*(6), 760–771. https://doi.org/10.1002/acp.1507

Carroll, M., Campbell-Ratcliffe, J., Murnane, H., & Perfect, T. (2007). Retrieval-induced forgetting in educational contexts: Monitoring, expertise, text integration, and test format. *European Journal of Cognitive Psychology*, *19*(4–5), 580–606. https://doi.org/10.1080/09541440701326071

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

Daniel, D. B., & Broida, J. (2004). Using web-based quizzing to improve exam performance: Lessons learned. *Teaching of Psychology*, *31*(3), 207–208. https://doi.org/10.1207/s15328023top3103_6

Dixon, P. (2008). Models of accuracy in repeated-measures designs. *Journal of Memory and Language*, *59*(4), 447–456. https://doi.org/10.1016/j.jml.2007.11.004

Downs, S. D. (2015). Testing in the college classroom: Do testing and feedback influence grades throughout an entire semester? *Scholarship of Teaching and Learning in Psychology*, *1*(2), 172–181. https://doi.org/10.1037/stl0000025

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

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

Foss, D. J., & Pirozzolo, J. W. (2017). Four semesters investigating frequency of testing, the testing effect, and transfer of training. *Journal of Educational Psychology*. https://doi.org/10.1037/edu0000197

Glass, A. L., Brill, G., & Ingate, M. (2008). Combined online and in-class pretesting improves exam performance in general psychology. *Educational Psychology*, *28*(5), 483–503. https://doi.org/10.1080/01443410701777280

Glover, J. A. (1989). The "testing" phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology*, *81*(3), 392–399.

Goossens, N. A. M. C., Camp, G., Verkoeijen, P. P. J. L., Tabbers, H. K., Bouwmeester, S., & Zwaan, R. A. (2016). Distributed practice and retrieval practice in primary school vocabulary learning: A multi-classroom study. *Applied Cognitive Psychology*, *30*(5), 700–712. https://doi.org/10.1002/acp.3245

Greving, S., Lenhard, W., & Richter, T. (2020). Adaptive retrieval practice with multiple-choice questions in the university classroom. *Journal of Computer Assisted Learning*. 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, 9:2412*. https://doi.org/10/gfkwvm

Halamish, V., & Bjork, R. A. (2011). When does testing enhance retention? A distribution-based interpretation of retrieval as a memory modifier. *Journal of Experimental Psychology. Learning, Memory & Cognition*, *37*(4), 801–812. https://doi.org/10.1037/a0023219

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

Ibrahim, J. G., & Molenberghs, G. (2009). Missing data methods in longitudinal studies: A review. *Test*, *18*(1), 1–43. https://doi.org/10.1007/s11749-009-0138-x

Jaeger, A., Eisenkraemer, R. E., & Stein, L. M. (2015). Test-enhanced learning in third-grade children. *Educational Psychology, 35*(4), 513–521. https://doi.org/10.1080/01443410.2014.963030

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*(4), 434–446. https://doi.org/10.1016/j.jml.2007.11.007

Jang, Y., Pashler, H., & Huber, D. E. (2014). Manipulations of choice familiarity in multiple-choice testing support a retrieval practice account of the testing effect. *Journal of Educational Psychology*, *106*(2), 435–447. https://doi.org/10.1037/a0035715

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

Kang, S. H. K., McDermott, K. B., & Roediger, H. L. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology*, *19*(4/5), 528–558. https://doi.org/10.1080/09541440601056620

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

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

Kornell, N., Bjork, R. A., & Garcia, M. A. (2011). Why tests appear to prevent forgetting: A distribution-based bifurcation model. *Journal of Memory and Language*, *65*(2), 85–97. https://doi.org/10.1016/j.jml.2011.04.002

Kromann, C. B., Jensen, M. L., & Ringsted, C. (2009). The effect of testing on skills learning. *Medical Education*, *43*(1), 21–27. https://doi.org/10.1111/j.1365-2923.2008.03245.x

Lahner, F.-M., Lörwald, A. C., Bauer, D., Nouns, Z. M., Krebs, R., Guttormsen, S., Fischer, M. R., & Huwendiek, S. (2018). Multiple true–false items: A comparison of scoring algorithms. *Advances in Health Sciences Education*, *23*(3), 455–463. https://doi.org/10/gfrd67

Larsen, D. P., Butler, A. C., & Roediger, H. L. (2009). Repeated testing improves long-term retention relative to repeated study: A randomised controlled trial. *Medical Education*, *43*(12), 1174–1181. https://doi.org/10.1111/j.1365-2923.2009.03518.x

Leeming, F. C. (2002). The exam-a-day procedure improves performance in psychology classes. *Teaching of Psychology*, *29*(3), 210–212. https://doi.org/10.1207/S15328023TOP2903_06

Leggett, J. M. I., Burt, J. S., & Carroll, A. (2019). Retrieval practice can improve classroom review despite low practice test performance. *Applied Cognitive Psychology*, acp.3517. https://doi.org/10/gf2rp2

Lipko-Speed, A., Dunlosky, J., & Rawson, K. A. (2014). Does testing with feedback help grade-school children learn key concepts in science? *Journal of Applied Research in Memory and Cognition, 3*(3), 171–176. https://doi.org/10.1016/j.jarmac.2014.04.002

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

Makowski, D., Ben-Shachar, M., & Lüdecke, D. (2019). bayestestR: Describing effects and their uncertainty, existence and significance within the Bayesian framework. *Journal of Open Source Software*, *4*(40), 1541. https://doi.org/10.21105/joss.01541

Marsh, E. J., Fazio, L. K., & Goswick, A. E. (2012). Memorial consequences of testing school-aged children. *Memory*, *20*(8), 899–906. https://doi.org/10.1080/09658211.2012.708757

Mayer, R. E., Stull, A., DeLeeuw, K., Almeroth, K., Bimber, B., Chun, D., Bulger, M., Campbell, J., Knight, A., & Zhang, H. (2009). Clickers in college classrooms: Fostering learning with questioning methods in large lecture classes. *Contemporary Educational Psychology*, *34*(1), 51–57. https://doi.org/10.1016/j.cedpsych.2008.04.002

McDaniel, M. A., Agarwal, P. K., Huelser, B. J., McDermott, K. B., & Roediger, H. L. (2011). Test-enhanced learning in a middle school science classroom: The effects of quiz frequency and placement. *Journal of Educational Psychology*, *103*(2), 399–414. https://doi.org/10.1037/a0021782

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

McDaniel, M. A., & Little, J. L. (2019). Multiple-choice and short-answer quizzing on equal footing in the classroom: Potential indirect effects of testing. In J. Dunlosky & K. A. Rawson (Eds.), *The Cambridge Handbook of Cognition and Education* (1st ed., pp. 480–499). Cambridge University Press. https://doi.org/10.1017/9781108235631.020

McDaniel, M. A., Thomas, R. C., Agarwal, P. K., McDermott, K. B., & Roediger, H. L. (2013). Quizzing in middle-school science: Successful transfer performance on classroom exams. *Applied Cognitive Psychology*, *27*(3), 360–372. https://doi.org/10.1002/acp.2914

McDaniel, M. A., Wildman, K. M., & Anderson, J. L. (2012). Using quizzes to enhance summative-assessment performance in a web-based class: An experimental study. *Journal of Applied Research in Memory and Cognition*, *1*(1), 18–26. https://doi.org/10.1016/j.jarmac.2011.10.001

McDermott, K. B., Agarwal, P. K., D'Antonio, L., Roediger, H. L., & McDaniel, M. A. (2014). Both multiple-choice and short-answer quizzes enhance later exam performance in middle and high school classes. *Journal of Experimental Psychology: Applied*, *20*(1), 3–21. https://doi.org/10.1037/xap0000004

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*, *4:5*. https://doi.org/10/gf2rp4

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

Palmer, S., Chu, Y., & Persky, A. M. (2019). Comparison of rewatching class recordings versus retrieval practice as post-lecture learning strategies. *American Journal of Pharmaceutical Education*, *83*(9). https://doi.org/10.5688/ajpe7217

Phelps, R. P. (2012). The effect of testing on student achievement, 1910–2010. *International Journal of Testing*, *12*(1), 21–43. https://doi.org/10.1080/15305058.2011.602920

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

Richter, T. (2006). What is wrong with ANOVA and multiple regression? Analyzing sentence reading times with hierarchical linear models. *Discourse Processes*, *41*(3), 221–250. https://doi.org/10.1207/s15326950dp4103_1

Roediger, H. L., Agarwal, P. K., McDaniel, M. A., & McDermott, K. B. (2011). Test-enhanced learning in the classroom: Long-term improvements from quizzing. *Journal of Experimental Psychology: Applied*, *17*(4), 382–395. https://doi.org/10.1037/a0026252

Roediger, H. L., & Karpicke, J. D. (2006a). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*(3), 181–210. https://doi.org/10.1111/j.1745-6916.2006.00012.x

Roediger, H. L., & Karpicke, J. D. (2006b). 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., & Marsh, E. J. (2005). The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(5), 1155–1159. https://doi.org/10.1037/0278-7393.31.5.1155

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* (Vol. 55, pp. 1–36). Elsevier. 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), 1432–1463. https://doi.org/10.1037/a0037559

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

Stenlund, T., Jönsson, F. U., & Jonsson, B. (2017). Group discussions and test-enhanced learning: Individual learning outcomes and personality characteristics. *Educational Psychology*, *37*(2), 145–156. https://doi.org/10.1080/01443410.2016.1143087

Stenlund, T., Sundström, A., & Jonsson, B. (2016). Effects of repeated testing on short- and long-term memory performance across different test formats. *Educational Psychology*, *36*(10), 1710–1727. https://doi.org/10.1080/01443410.2014.953037

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

Verbic, S. (2012). Information value of multiple response questions. *Psihologija*, *45*(4), 467–485. https://doi.org/10/gfrd64

Vojdanoska, M., Cranney, J., & Newell, B. R. (2010). The testing effect: The role of feedback and collaboration in a tertiary classroom setting. *Applied Cognitive Psychology*, *24*(8), 1183–1195. https://doi.org/10.1002/acp.1630

Weinstein, Y., Nunes, L. D., & Karpicke, J. D. (2016). On the placement of practice questions during study. *Journal of Experimental Psychology: Applied*, *22*(1), 72–84. https://doi.org/10.1037/xap0000071

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

Yong, P. Z., & Lim, S. W. H. (2016). Observing the testing effect using coursera video-recorded lectures: A preliminary study. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.02064

Zhang, M., Chen, X., & Liu, X. L. (2019). Confidence in accuracy moderates the benefits of retrieval practice. *Memory*, *27*(4), 548–554. https://doi.org/10/gfwwj8

Chapter IV

The Testing Effect in University Teaching: Using Multiple-Choice Testing to Promote Retention of Highly Retrievable Information

Study 3

A version of this chapter was published as:

Greving, S., Lenhard, W., & Richter, T. (2022). The testing effect in university teaching: Using multiple-choice testing to promote retention of highly retrievable information. *Teaching of Psychology* (online first). https://doi.org/10.1177/00986283211061204

The Testing Effect in University Teaching: Using Multiple-Choice Testing to Promote Retention of Highly Retrievable Information

Sven Greving, Wolfgang Lenhard, & Tobias Richter

Abstract. Retrieval practice promotes retention of learned information more than restudying the information. However, benefits of multiple-choice testing over restudying in real-world educational contexts and the role of practically relevant moderators such as feedback and learners' ability to retrieve tested content from memory (i.e., retrievability) are still underexplored. The present research examines the benefits of multiple-choice questions with an experimental design that maximizes internal validity, while investigating the role of feedback and retrievability in an authentic educational setting of a university psychology course. After course sessions, students answered multiple-choice questions or restudied course content and afterwards could choose to revisit learning content and obtain feedback in a self-regulated way. Participants obtained corrective feedback only for 2% of practiced items when practicing course content. In the criterial test, practicing retrieval was not superior to reading summarizing statements in general but a testing effect emerged for questions that targeted information that participants could easily retrieve from memory. Feedback was rarely sought. However, even without feedback, participants profited from multiple-choice questions that targeted easily retrievable information. Caution is advised when employing multiple-choice testing in self-regulated learning environments in which students are required to actively obtain feedback.

Students often use (electronic) flashcards, and teachers use clicker questions to encourage the practice of learning content (Caldwell, 2007; Chauhan, 2017; Golding et al., 2012; Mayer et al., 2009; Wissman et al., 2012). Digital flashcards and online quizzes have the common purpose of responding to questions about the learning content. Learners using these technologies knowingly or unknowingly benefit from the testing effect, also known as retrieval practice effect. The testing effect means that active recall from practicing learned content is more beneficial for retention than restudying the same content. This testing effect has been reliably found in many laboratory studies (c.f., the meta-analyses by Adesope et al., 2017; Phelps, 2012; Rowland, 2014). Consequently, research has focused on the application of the testing effect in real-world educational contexts (Adesope et al., 2017; meta-analysis by Schwieren et al., 2017). However, the role of feedback for benefits of multiple-choice testing over restudying in applied contexts is still largely unclear (McDaniel & Little, 2019; Moreira et al., 2019; Roediger et al., 2011). Applied research has indicated that whenever feedback is withheld, multiple-choice testing is not more effective than restudying (Greving & Richter, 2018). Feedback seems especially important for students who engage in self-regulated restudying after unsuccessful testing. The aim of this article is to examine the testing effect for multiple-choice questions in an authentic educational setting of a university course while also allowing students to obtain corrective feedback in a naturalistic way.

Testing Effects in Laboratory and Applied Research

The effectiveness of answering multiple-choice questions for promoting retention of learned content has been demonstrated in many laboratory studies. Summarizing this research, meta-analyses have found small to medium effect sizes (Hedges' g) ranging between 0.36 and 0.70 (Adesope et al., 2017; Rowland, 2014). These analyses also revealed that the most important moderators of the testing effect are retrievability (Rowland, 2014) and feedback (Adesope et al., 2017; Phelps, 2012; Rowland, 2014). Retrievability in this

context describes the success with which learning content can be retrieved from memory, resulting in correct responses in the testing condition. Therefore, retrievability can be operationalized by the difficulty of items in the practice tests. In contrast, corrective feedback seems to increase the testing effect mainly by fostering learning from unsuccessful retrieval attempts (Pashler et al., 2005).

Providing learners with multiple-choice questions instead of restudying opportunities is easy and, therefore, bears great potential for its application in educational settings. To determine the effectiveness of multiple-choice testing for practical purposes, applied research should provide evidence that these tests lead to better retention than other revision strategies such as restudying. Furthermore, applied research should provide evidence for whether factors found to influence the testing effect in laboratory settings are equally significant in educational contexts. These results can inform practitioners in developing guidelines on when and how to use multiple-choice testing as a tool for fostering retention. Proponents of the testing effect assume a beneficial effect of testing compared to restudying, irrespective of feedback (e.g., Roediger & Karpicke, 2006). However, a recent review of testing in educational contexts identified that all studies on multiple-choice testing had employed feedback or had not compared testing to another activity (Moreira et al., 2019). Consequently, the authors stated that no conclusions can be drawn on whether multiple-choice testing alone is superior to other activities that foster retention. This conclusion is especially worrying since recent studies have not found multiple-choice testing superior to restudying when simulating classroom learning under laboratory conditions (Butler & Roediger, 2007; Kang et al., 2007; Nungester & Duchastel, 1982). Additionally, a study that investigated multiple-choice testing without feedback in a university lecture found that

restudying was equally effective as taking multiple-choice quizzes and that retrievability of practice questions had no effect on these outcomes (Greving & Richter, 2018).

The differences in findings of studies that have investigated multiple-choice testing with and without feedback can be partly attributed to various indirect factors in addition to benefits of testing on memory (McDaniel & Little, 2019; Roediger et al., 2011). Consequently, additional research is needed that investigates the testing effect of multiple-choice questions and its potential moderators in educational contexts.

Feedback and the Testing Effect

Feedback is an important component in self-regulated learning that shapes cognitive and metacognitive processes alike: It can confirm information and beliefs students hold, correct erroneous information or beliefs, add knowledge, tune motivation and beliefs, or restructure knowledge (Butler & Winne, 1995). When students learn in a self-regulated fashion, feedback can arise from various sources such as self-generation or external sources (Bangert-Drowns et al., 1991). Feedback also comes in different types such as formative or summative and focusses on different aspects of learning such as learning outcomes, meta-memory, or motivation (Butler & Winne, 1995). In research on the testing effect, it is most common to use feedback that is formative, external and outcome-oriented and indicates whether a given answer is correct, with incorrect answers often followed by the correct answer. Feedback following practice tests provides two kinds of information to students. For one, it helps correct erroneous information, confirm correct information or add knowledge and can thus exert a memorial effect (Pashler et al., 2005). Furthermore, it can provide a more realistic judgement of the learning outcomes and guide future learning activities and thus operate on a metacognitive level (e.g., Kornell & Rhodes, 2013).

Examining the influence of feedback in real-world educational settings and whether it adds to the unmediated effects of practice tests is important for several reasons. First, very few applied studies have investigated the effects of feedback on practice testing, and these studies have yielded mixed results. Two studies found beneficial effects of feedback on the testing effect (Marsh et al., 2012; Vojdanoska et al., 2010). However, among other shortcomings, both studies lacked a restudy control condition. Additionally, Butler and Roediger (2007) found no effects of feedback on the testing effect. However, the study was conducted in a simulated classroom and the effect of feedback given in this context might have differed from the effect of feedback in real-world educational contexts.

Second, studies investigating testing effects in educational contexts without the provision of corrective feedback might have suffered from limited ecological validity. We assume that students gain an additional metamemorial benefit from practice testing as they become aware of their knowledge gaps. This assumption is backed by findings that practice tests lead to more accurate memory predictions (Little & McDaniel, 2015), which in turn influence students' decisions for or against additional studying (Metcalfe & Finn, 2008). Consequently, students who use practice testing compared to students who restudy might therefore engage in additional restudying to close knowledge gaps that become apparent from practicing questions. Thus, research is necessary that assesses students' need for feedback following different practice opportunities.

Third, lab studies have shown beneficial effects of feedback when practicing multiple-choice questions, but this effect can also avert the negative effects that could arise because of the exposure to incorrect information in the form of lures or distractors (Butler & Roediger, 2008; Roediger & Marsh, 2005). Therefore, research is needed to investigate

whether feedback adds to the unmediated effects of multiple-choice practice tests in educational settings.

Rationale of This Study

The present study adds to the body of research by investigating the beneficial effects of multiple-choice questions without feedback compared to restudying in an educational setting. Additionally, we incorporated two moderators that might be crucial for the effectiveness of multiple-choice practice tests in a real-world educational context. The first potential moderator is retrievability of learned content when the learned content is practiced. The second potential moderator is the possibility to obtain feedback for responses in the practice test. In applied research that investigated the beneficial effects of multiple-choice testing with additional feedback, researchers advocating this procedure have implicitly assumed that feedback will always be provided by instructors or that students will seek feedback whenever they do not know the answer to practice questions. Questioning this assumption, we employed a more naturalistic manipulation of feedback. Participants were allowed to revisit the learning material whenever needed while practicing the learned content. This method is similar to the type of feedback found in educational settings that incorporate self-regulated learning.

We manipulated students' follow-up learning opportunities after they visited a university course session. Practice consisted of answering multiple-choice questions about the learning content or restudying the learning content. Students could revisit the original content after practicing to receive corrective feedback. We assessed the effectiveness of practice testing vs. restudying with a surprise criterial test administered between 1 and 7 days after the last practice session. The original learning materials were textbook chapters that were well written, clearly structured, and did not require a large amount of specific prior

knowledge. Thus, the complexity (or element interactivity) of the learning materials in terms of cognitive load theory was rather low, which should create favorable conditions for the testing effect to occur (Hanham et al., 2017; Van Gog & Sweller, 2015).

Prior research suggests that being tested on learned information is associated with a critical evaluation of one's knowledge of the learning content (Metcalfe & Finn, 2008).

Given that there is considerable evidence for testing effects in real-world educational settings (Adesope et al., 2017; Schwieren et al., 2017), we expected a testing effect. From research investigating the retrievability of tested items (Greving & Richter, 2018) and beneficial effects of feedback on the testing effect, we assumed both retrievability and feedback to add to the testing effect.

We tested the following four operational hypotheses derived from our assumptions: Hypothesis 1 states that answering multiple-choice questions leads to more request for feedback in terms of revisiting the original content compared to restudying. Hypothesis 2 states that in terms of retention of the learning content, the testing condition outperforms the restudy condition. Hypotheses 3 and 4 state that higher retrievability rates and more corrective feedback are associated with more pronounced testing effects, respectively.

Method

Participants, Power, and Required Sample Size

We aimed to recruit participants from university courses on the topic of behavioral and learning disorders in childhood and youth. Usually, 35 students participate in each of these courses. The students in these courses were enrolled in a teacher training program and would eventually become teachers in different school tracks. The courses belong to the mandatory psychology curriculum of the study program. Students could freely choose between eight different seminars and one lecture, all of which conveyed the same content and

prepared for the same exam. The final exam was scored automatically with a fixed scoring rubric.

We did not know beforehand how many of the students in a course would participate in the study. Therefore, we simulated the statistical power necessary to detect a testing effect that was similar to those reported in meta-analyses investigating the testing effect with feedback in the psychology classroom (Schwieren et al., 2017). In these power analyses we simulated the outcomes of multiple generalized linear mixed models (GLMM, see Results section) that matched our planned analyses for different sample sizes and checked whether the confidence interval of a simulated statistical power surpassed the 80% threshold (Brysbaert & Stevens, 2018). Model parameters were taken from a study that investigated the testing effect for multiple-choice tests and differing retrievability in the university classroom (Greving & Richter, 2018). Power simulations revealed that planned analyses with only 20 students would not reliably provide enough statistical power within a 95% confidence interval (power = 81%, CF [79, 83]), however 25 (power = 87%, CF [85, 89]), 30 (power = 89%, CF [87, 91]), or 40 (power = 97%, CF [95, 98]) students would be sufficient. Based on this power analysis and considering that not all students would opt to participate in the study, we chose to recruit students from two courses to make sure that the minimum number of 25 participants is reached.

Participants gave their informed and written consent prior to participation and they received course credit for participating in form of a small bonus on their final course grade regardless of their performance in the experiment. Additionally, students were informed that participation in this study involved practicing the learning content taught in these courses and thus was assumed to yield better preparation for the upcoming exam. The decision not to participate in the study did not yield any negative consequences. Moreover, there were other

opportunities to earn the bonus on the final course grade. Given the flexibility of exam choice and the objective automatic scoring of the results, students could be assured that their participation had no negative effect on their grades. The study was in accordance with all ethical guidelines of the German Psychological Association. According to these guidelines, explicit approval of the ethics committee of the Institute of Psychology was not required.

Thirty students (73% female) completed all parts of the study. Participants' age ranged from 18 to 24 years (M = 19.80, SD = 1.49), and participants were mostly students in their second semester (M = 2.13, SD = 0.68) of their study program, whose length ranges from seven (elementary school) to nine (upper educational tracks and special education) semesters. The courses in psychology are recommended for students to attend early in the first year of their studies, because the contents taught in the psychology courses partly provide the basis for other courses in the teacher curriculum. However, the recommendations are not binding, and students can deviate from them in their study plan.

Materials

Test Items and Restudy Statements

Two chapters from a textbook on mental disorders that are part of the regular reading assignments of the course were selected as the basis for study material. We identified text segments that reflected the logical structure of the text and represented one key information unit of the text. The content of the chapter on "Drug abuse and addiction" was divided into 58 text segments and the chapter on "Suicidality" was divided into 37 text segments. Text segment length ranged between 28 and 255 words (M =113.21, SD = 46.64). Each text segment was made available for revisiting as part of the feedback. For 20 information units from each chapter, one statement and one multiple-choice question for each unit were created by summarizing the key information of the information unit. An example

statement is: “The prevalence of alcohol addiction in adolescents is identical to the prevalence of alcohol addiction in the total population: 2–3%. Girls make up one fifth of the juvenile alcohol addicts.” Multiple-choice questions with four response options were created by asking for the key information, for example: “Which statement concerning the prevalence of addiction to alcohol is correct? (A, correct option) Girls make up a fifth of the juvenile alcohol addicts, (B) Girls make up a third of the juvenile alcohol addicts, (C) The prevalence of alcohol addiction in the total population is 15%, (D) The prevalence of alcohol addiction among adolescents is 15%.”

Practice Materials and Feedback

Practice materials and feedback were presented with the software Inquisit (Version 5.0.6.0; Millisecond Software, 2016) and consisted of either 20 multiple-choice questions (testing condition) or 20 summarizing statements (restudy condition). In both conditions, each practice item was presented on one page.

Whenever participants requested feedback, they were first presented with a table of contents from which to choose text segments to revisit. After choosing a text segment, the segment was displayed and participants were free to navigate the text by using arrow keys or by returning to the table of contents. Text segments served as the basis for the initial creation of idea units. Conversely, the information of each question and answer and each statement could be found in the corresponding text segment.

The experimental software recorded which text segments were revisited (if any) for each preceding practice item.

Criterial Test

A criterial test was constructed that consisted of 14 items from each topic, resulting in 28 questions. All items were identical to the practice items in the testing condition.

Design

We investigated the effect of the independent variable practice condition (testing or restudy) across two course sessions on the dependent variable performance in the criterial test. All participants experienced all practice conditions in the course sessions (within-participants design). To control for effects of topic and sequence, we counterbalanced the sequence of conditions. Each participant was randomly assigned to one of the two sequences upon arrival at the first practice session.

Procedure

General Procedure

Students in the course were assigned to read book chapters in preparation for the course sessions. All course sessions were taught by the first author, and the course content was based on and followed the thematic structure of the reading assignments but contained further explanations and details for a deeper understanding of the contents. All content addressed in the upcoming exam questions occurred in both the literature and the course. The study was conducted in the first four weeks of the semester whereas after two subsequent regular course sessions (i.e., the focal sessions), voluntary practice sessions were offered, in which the experiment was conducted (see Figure IV.1). In Week 1, students were informed about the study and the course and given the first reading assignment in preparation of the second course session. In Week 2, the course session on “Suicidality” took place (first focal session). At the end of the course session, students were given the second reading assignment in preparation of the third course session and study participants were asked to practice the

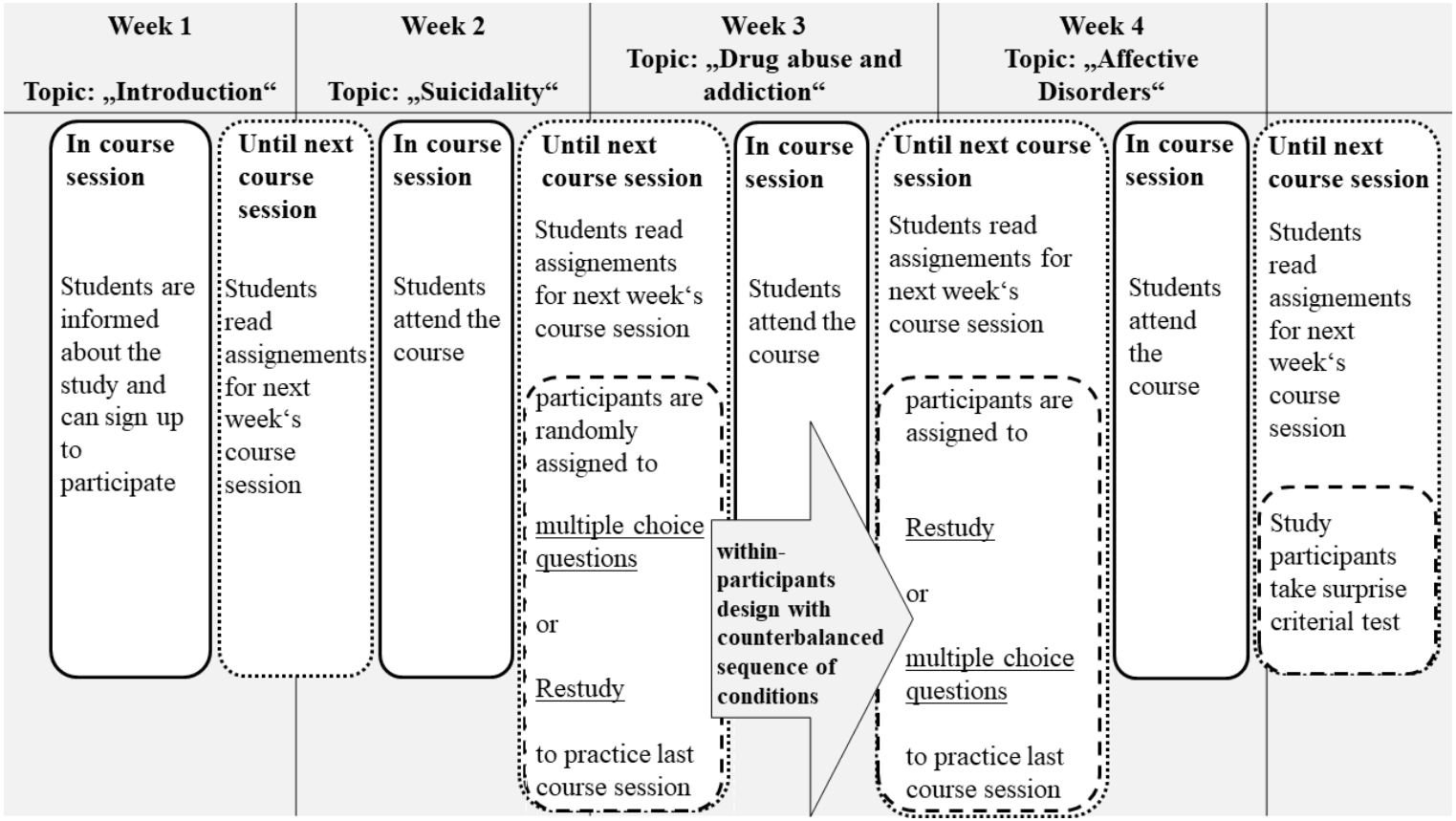
course content of the last session in the laboratory within 1 week. In Week 3, the course session on “Drug abuse and addiction” took place (second focal session). At the end of the course session, study participants were asked to practice the course content of the last session in the laboratory within 1 week. In Week 4, the next course session took place that was unrelated to the study and at the end of the session, study participants were instructed to present themselves to the laboratory within 1 week, which was announced as an additional practice session. In this session in the lab, the surprise criterial test was administered.

Practice Sessions and Criterial Test

Practice sessions were conducted in the laboratory and took between 15 and 50 min, depending on participants’ speed and willingness to obtain feedback. In each practice session, participants first answered sociodemographic items and reported whether they fulfilled the reading assignment. Participants then engaged in practicing the course content in one of the two practice conditions (testing or restudy). Practice was self-paced. The order of practice items was randomized. In the restudy condition, participants were asked to read statements, whereas in the testing condition, participants were asked to answer multiple-choice questions. In the restudy condition, participants were then asked whether they wanted to revisit the original learning content. In the testing condition, a message first indicated whether an item was answered correctly before asking the participants whether they wanted to revisit the original learning content. Whenever participants gave an affirmative response in both conditions, they were allowed access to the original learning content and to browse and reread it with no time limitation.

Figure IV.1

Visual Presentation of Design and General Procedure by Week



Note. In course activities are displayed differently (solid outline) than activities between course sessions (dotted outline). Additionally, study participants' activities are furthermore visually differentiated from student activities (dashed outline).

Results

Learning Behavior, Retrieval Practice Performance, and Feedback Seeking Behavior

Seventeen participants reported having read the chapter on “Suicidality” and 15 participants indicated having read the chapter on “Drug abuse and addiction”. Participants in the practice condition answered between 3 and 16 of the 20 items correctly, on average 51% (number of correct answers: M = 10.10, SD = 3.37). A majority (54%) answered more than half of the questions correctly. Participants requested feedback in the form of revisiting the text for 0–9 items in the practice sessions, on average for only 9% (number of items: M = 3.53, SD = 5.83). We investigated whether the practice condition had an effect on feedback behavior (Hypothesis 1): Participants requested feedback following items in the testing condition (M = 2.07, SD = 1.36) slightly more often than for items in the restudy condition (M = 1.47, SD = 1.94), but the difference was not significant, $t(51.98) = 1.38$, $p = .172$, $d = 0.36$, CI [-0.27, 1.47]. Of 106 total revisits of the text, only 19 (18%) text segments included the content of the preceding practice item.

In sum, the practice items in the testing condition were relatively difficult. Nevertheless, overall request for feedback was small, the requests did not differ between conditions, and the relevant information was mostly not found in the learning materials. We therefore concluded that participants feedback behavior was too sporadic and erratic to be considered appropriate in terms of corrective feedback.

Modeling Performance in the Criterial Tests

Hypotheses 2, 3, and 4 assume effects of different variables on retention of course content as measured in the criterial test. These variables varied between items (retrievability and practice condition) or as a function of participants and items (feedback behavior of

participants for items). Additionally, considering that—similar to real educational setting—participants varied in their ability to retain content and learning content varied in its memorability. We chose to adapt multi-level modeling because this approach allows for simultaneous investigation of predictors on the item level (retrievability and practice condition) and on the combined participant-item level (feedback behavior per item) while controlling for unsystematic variance on both the item and participant level. Mixed effect models have many advantages compared to ANOVAs (e.g., Baayen et al., 2008; Richter, 2006) especially with regard to analyzing categorical outcome variables (Jaeger, 2008). We estimated GLMMs with a logit-link function (Dixon, 2008) with the R package lme4 (Bates et al., 2015). We used the package emmeans (Lenth, 2019) for comparisons between experimental conditions and for estimating performance scores for different conditions. Participants and test items were included as random effects (random intercepts) in all models.

The testing condition was compared to the restudy condition (dummy coded: testing = 1, restudy = 0).

We included the retrievability of learned information with two dummy-coded predictors that contrasted items of medium retrievability and low retrievability with items of high retrievability as the reference condition. To construct this predictor, we grouped the multiple-choice questions into three equally sized ordered categories (tertiles) according to their difficulty in the practice tests. To avoid distortions from extreme values, we discarded the lowest and the highest 5% of the distribution before the grouping, which led to the exclusion of three items. Item difficulties were corrected for guessing by subtracting the times an item was answered incorrectly from the times an item was answered correctly and dividing this difference by the available multiple-choice options - 1 (e.g., Frary, 1988). This procedure resulted in three categories of items with high (item difficulties from 56% to 91%),

medium (21%–55%), or low retrievability (0%–20%). Following this rationale, retrievability of 10 of the items were classified as being easily retrievable, whereas 11 and 7 items were associated with medium and low retrievability, respectively. Finally, the models included the interaction of retrievability with testing vs. restudying. We refrained from including feedback behavior as a predictor because of the sporadic and erratic way participants used feedback. All predictors and the interactions were entered simultaneously in the models.

To account for the possibility of the length of the text section exerting direct or indirect effects on retention of the learning content, we also estimated a model that included the z-transformed text length as well as its interactions with all other predictors. However, model fit for this model was not significantly better than for the more parsimonious original model $\Delta\chi2(6) = 5.48$, $p = 0.484$. Thus, the model including text length was discarded.

Model-based Hypothesis Testing

As Hypotheses 2, 3, and 4 assume an influence of certain variables on the retention of learning content as measured in the criterial test, determining the significance of factors in the estimated model serves as hypotheses tests. As the erratic nature of the feedback behavior led us to exclude feedback behavior, Hypothesis 4 could not be tested.

The model estimates for the original model are shown in Table 1. In terms of hypotheses 2 and 3, a positive effect for testing emerged $\beta = 0.63$, $SE = 0.32$, $p = .025$, one-tailed, $OR = 1.88$, 95% CI [1.00, 3.5]. However, negative interactions of the practice condition with the predictor comparing low to high retrievability $\beta = -0.93$, $SE = 0.46$, $p = .045$, $OR = 0.40$, 95% CI [0.16, 0.98] and the interaction with the predictor comparing medium to high retrievability were significant $\beta = -0.98$, $SE = 0.39$, $p = .012$, $OR = 0.40$, 95% CI [0.17, 0.81]. Resulting mean probabilities of correct responses in the criterial test are depicted in Figure 2. Planned contrasts revealed a testing effect only for items with high

retrievability z = 1.97, p = .024, one-tailed, OR = 1.88, 95% CI [1.00, 3.53] but not for items with medium z = -1.52, p = .936, one-tailed, OR = 0.70, 95% CI [0.44, 1.11] or low retrievability z = -0.94, p = .825, OR = 0.74, 95% CI [0.40, 1.39]. Table IV.2

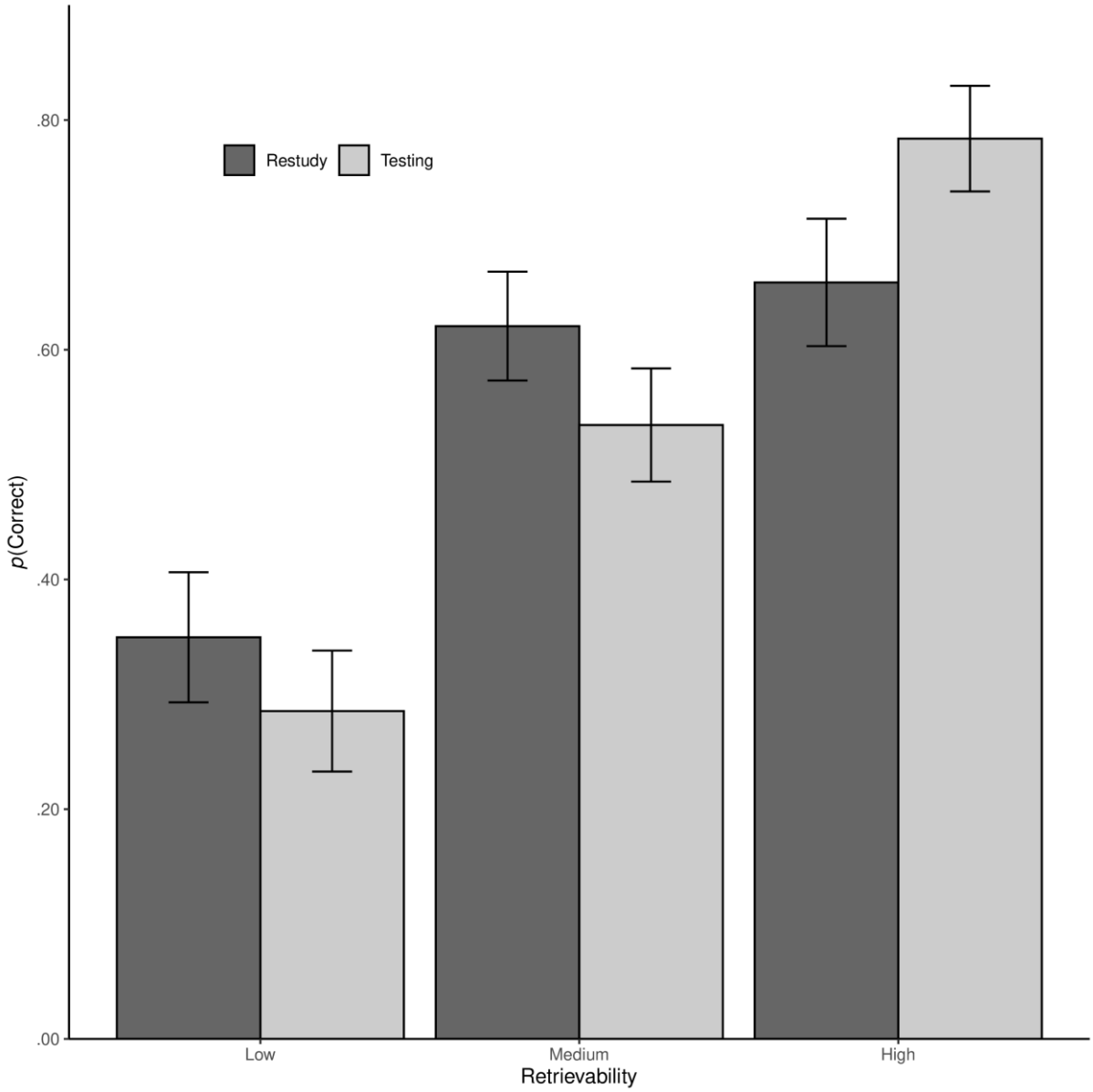
Model Parameters

Parameter	β	*SE*	*z*	*p*
Intercept	0.66	0.25	2.67	.008
Testing	0.63	0.32	1.96	.025 [a]
Low retrievability	-1.28	0.33	-3.86	<.001
Medium retrievability	-0.17	0.29	-0.57	.567
Testing x Low retrievability	-0.93	0.46	-2.00	.977 [a]
Testing x Medium retrievability	-0.98	0.39	-2.50	.994 [a]
$N_{Participants}$			30	
N_{Items}			25	

Note. Parameter estimates for a model estimating the effects of testing and retrievability on criterial test performance. Testing (dummy coded: testing = 1, restudy = 0). Low retrievability (dummy coded: low retrievability = 1, high retrievability = 0). Medium retrievability (dummy coded: medium retrievability = 1, high retrievability = 0). [a]p values refer to one-tailed tests for $\beta > 0$. Other *p* values refer to two-tailed tests.

Figure IV.1

Probability of a Correct Response in the Criterial Test by Retrievability and Practice Condition



Note. Mean probability of correct responses (with standard errors) in criterial test items (back-transformed from the logits in the GLMM) by retrievability (low, medium, or high), and practice condition (testing vs. Restudy).

Discussion

We investigated the effects of multiple-choice testing compared to restudying within an existing university course using a minimal intervention design. We expected the benefit of

testing to be dependent on the retrievability of the tested items and the amount of feedback learners were able to obtain by revisiting the learning content.

One main finding was that learners were unwilling to obtain feedback by revisiting the text, and whenever they revisited the text, they were often unable to identify relevant text segments that corresponded to the practiced content. Contrary to our assumptions, feedback requests were independent of practice condition (Hypothesis 1). The main consequence of this feedback behavior was that only 2% percent of practiced items were followed by feedback that included the correct information. We therefore assume that no corrective feedback was provided in this study and thus interpret the remaining results accordingly and refrain from testing Hypothesis 4. It is beyond the scope of this study to investigate reasons for the unwillingness to obtain feedback, however it is reasonable to assume a variety of factors that might influence feedback on a general level, such as study participants shying the additional effort. It is a strength of this study to investigate the effects of multiple-choice testing in a naturalistic setting while using course material that eventually was also tested in the exam, and that participants were informed that practicing course content (which also includes multiple-choice questions with additional feedback) will help them prepare for the upcoming exam. The naturalistic setting suggests that the observed feedback behavior observed indeed reflects students' willingness to engage in more learning subsequent to being tested. Additional research, however, should investigate participant-specific explanations for the lack of feedback behavior such as meta-cognitive illusions (Metcalfe & Finn, 2008) or experienced stereotype threat imposed by the testing situation (e.g., Mangels et al., 2012).

Another main finding was that the beneficial effects of testing largely depended on the retrievability of the practiced items (Hypothesis 3): Contrary to Hypothesis 2, we found a testing effect for highly retrievable items only, which corresponds to mean retrievability rates

between 56% to 91%. To our knowledge, this result is the first evidence for a testing effect for multiple-choice questions in a real-world educational context, without the provision of corrective feedback and in comparison to another activity that fosters retention.

Beneficial effects of practicing highly retrievable multiple-choice questions without feedback are in line with the bifurcation model (Halamish & Bjork, 2011; Kornell et al., 2011), with previous findings from laboratory research (Rowland, 2014), and also findings from research in educational contexts using short-answer questions (Butler & Roediger, 2007; Greving & Richter, 2018). The bifurcation model states that the superiority of testing without feedback compared to restudying is largely dependent on the amount of successfully retrieved items in the testing condition. This assumption can be partly backed up by the findings from Rowland's meta-analysis, which shows no testing effects for laboratory studies with no corrective feedback and retrievability rates of less than or equal to 50%. Similar interaction patterns were found in studies investigating the testing effect in educational settings. In these studies, a testing effect emerged for short-answer questions for which retrievability rates were higher than 50% (Butler & Roediger, 2007; Greving & Richter, 2018). Note that these studies followed a rationale similar to the present study. However, both studies found no testing effects for highly retrievable multiple-choice questions. This discrepancy of results might be explained by the delay of practice tests and the resulting increase in test difficulty. In the two previous studies, participants answered multiple-choice questions immediately after the initial study, whereas in the present study, time between initial study and practice conditions ranged between 1 day and 1 week. Research has shown that delaying an initial retrieval attempt increases retrieval difficulty, which in turn promotes long-term retention, provided that the retrieval is successful (Karpicke & Roediger, 2007). This explanation is in line with theoretical accounts stating that more difficult practice leads

to better long-term retention (Bjork, 1994; Pyc & Rawson, 2009) and with findings from studies in university classrooms (Greving et al., 2020).

We demonstrated in this study that multiple-choice testing increased the retention of learned content, whenever the learned content was retrievable. Learners had little need for corrective feedback. We emphasize the ecological validity of the current study because the method was implemented in an existing university course, and feedback was operationalized in a naturalistic way that closely resembles how students usually obtain feedback in self-regulated learning at the university.

Although the field-experimental approach is a strength of the study in terms of ecological validity it also presents some limitations. Compared to laboratory experiments, external influences potentially play a much greater role in a field setting, which are unknown to researchers and difficult to control. This limitation applies especially to students studying behavior between recorded practice and tests and also to metamemorial, motivational, and metacognitive variables. Another general limitation is that we examined the effectiveness of practice tests with multiple-choice items in just a single course with one sample of participants. The extent that the results generalize to student populations in higher education in general remains a question to be clarified in further research. A third limitation is that the practice and criterial-test questions used in this experiment were fact-based questions that drew on information explicitly provided in the learning materials or easily inferred. Especially in higher education, the ultimate goal is to teach transferable knowledge that can be applied to novel questions and problems. Our results cannot answer the question whether retrieval practice, even without feedback, enhances learning transfer. However, remembering the to-be-learned information may be seen as a necessary although not sufficient condition for transfer. Moreover, based on the meta-analysis by Pan and Rickard (2018), one may

speculate that retrieval practice in a university classroom, as implemented in our study, also has the potential to increase performance on typical transfer questions, such as application and inference questions.

To conclude, in this research we were able to demonstrate a testing effect for multiple-choice questions in a real-world educational context while using an experimental design that minimized common methodological issues. We also demonstrated that research that has identified retrievability as an important factor to consider when practicing short-answer questions in university teaching can be extended to multiple-choice question. In real-world educational settings, practitioners can exert influence on the retrievability by enabling students to answer practice questions correctly. However, practitioners should proceed with caution when employing multiple-choice testing in self-regulated learning environments in which students are required to actively obtain feedback. To profit the most from the testing effect, practitioners and textbook authors should thus identify the most important aspects of the learning content and foster a clear understanding of these prior to practicing multiple-choice tests.

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

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1). https://doi.org/10.18637/jss.v067.i01

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). MIT Press.

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

Butler, A. C., & Roediger, H. L. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition, 36*(3), 604–616. https://doi.org/10.3758/MC.36.3.604

Caldwell, J. E. (2007). Clickers in the large classroom: Current research and best-practice tips. *CBE—Life Sciences Education, 6*(1), 9–20. https://doi.org/10.1187/cbe.06-12-0205

Chauhan, J. (2017). Quiz in MOOC: An overview. *International Research Journal of Engineering and Technology (IRJET), 4*(7), 303–307.

Dixon, P. (2008). Models of accuracy in repeated-measures designs. *Journal of Memory and Language, 59*(4), 447–456. https://doi.org/10.1016/j.jml.2007.11.004

Golding, J. M., Wasarhaley, N. E., & Fletcher, B. (2012). The use of flashcards in an introduction to psychology class. *Teaching of Psychology, 39*(3), 199–202. https://doi.org/10.1177/0098628312450436

Greving, S., Lenhard, W., & Richter, T. (2020). Adaptive retrieval practice with multiple-choice questions in the university classroom. *Journal of Computer Assisted Learning, 36*(6), 799–809. 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, 9:2412*. https://doi.org/10/gfkwvm

Halamish, V., & Bjork, R. A. (2011). When does testing enhance retention? A distribution-based interpretation of retrieval as a memory modifier. *Journal of Experimental Psychology. Learning, Memory & Cognition, 37*(4), 801–812. https://doi.org/10.1037/a0023219

Kang, S. H. K., McDermott, K. B., & Roediger, H. L. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology, 19*(4/5), 528–558. https://doi.org/10.1080/09541440601056620

Karpicke, J. D., & Roediger, H. L. (2007). Expanding retrieval practice promotes short-term retention, but equally spaced retrieval enhances long-term retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*(4), 704–719. https://doi.org/10.1037/0278-7393.33.4.704

Kornell, N., Bjork, R. A., & Garcia, M. A. (2011). Why tests appear to prevent forgetting: A distribution-based bifurcation model. *Journal of Memory and Language, 65*(2), 85–97. https://doi.org/10.1016/j.jml.2011.04.002

Little, J. L., & McDaniel, M. A. (2015). Metamemory monitoring and control following retrieval practice for text. *Memory & Cognition, 43*(1), 85–98. https://doi.org/10.3758/s13421-014-0453-7

Marsh, E. J., Fazio, L. K., & Goswick, A. E. (2012). Memorial consequences of testing school-aged children. *Memory, 20*(8), 899–906. https://doi.org/10.1080/09658211.2012.708757

Mayer, R. E., Stull, A., DeLeeuw, K., Almeroth, K., Bimber, B., Chun, D., Bulger, M., Campbell, J., Knight, A., & Zhang, H. (2009). Clickers in college classrooms: Fostering learning with questioning methods in large lecture classes. *Contemporary Educational Psychology, 34*(1), 51–57. https://doi.org/10.1016/j.cedpsych.2008.04.002

McDaniel, M. A., & Little, J. L. (2019). Multiple-choice and short-answer quizzing on equal footing in the classroom: Potential indirect effects of testing. In J. Dunlosky & K. A. Rawson (Eds.), *The Cambridge Handbook of Cognition and Education* (1st ed., pp. 480–499). Cambridge University Press. https://doi.org/10.1017/9781108235631.020

Metcalfe, J., & Finn, B. (2008). Evidence that judgments of learning are causally related to study choice. *Psychonomic Bulletin & Review, 15*(1), 174–179. https://doi.org/10.3758/PBR.15.1.174

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, 4:5*. https://doi.org/10/gf2rp4

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

Pashler, H., Cepeda, N. J., Wixted, J. T., & Rohrer, D. (2005). When does feedback facilitate learning of words? J*ournal of Experimental Psychology: Learning, Memory, and Cognition, 31*(1), 3–8. https://doi.org/10.1037/0278-7393.31.1.3

Phelps, R. P. (2012). The effect of testing on student achievement, 1910–2010. *International Journal of Testing, 12*(1), 21–43. https://doi.org/10.1080/15305058.2011.602920

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

Richter, T. (2006). What is wrong with ANOVA and multiple regression? Analyzing sentence reading times with hierarchical linear models. *Discourse Processes, 41*(3), 221–250. https://doi.org/10.1207/s15326950dp4103_1

Roediger, H. L., & Karpicke, J. D. (2006). 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., & Marsh, E. J. (2005). The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(5), 1155–1159. https://doi.org/10.1037/0278-7393.31.5.1155

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* (Vol. 55, pp. 1–36). Elsevier. 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), 1432–1463. https://doi.org/10.1037/a0037559

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

Vojdanoska, M., Cranney, J., & Newell, B. R. (2010). The testing effect: The role of feedback and collaboration in a tertiary classroom setting. *Applied Cognitive Psychology, 24*(8), 1183–1195. https://doi.org/10.1002/acp.1630

Chapter V

Adaptive Retrieval Practice with Multiple-Choice Questions in the University Classroom

Study 4

A version of this chapter was published as:

Greving, S., Lenhard, W., & Richter, T. (2020). Adaptive retrieval practice with multiple-choice questions in the university classroom. *Journal of Computer Assisted Learning*, *36*(6), 799–809. https://doi.org/10.1111/jcal.12445

Adaptive Retrieval Practice with Multiple-Choice Questions in the University Classroom

Sven Greving, Wolfgang Lenhard, & Tobias Richter

Abstract. Retrieval practice has been shown to promote retention of learned information more than restudying the information (i.e., the testing effect) and is applied to many educational settings. However, little research has investigated means to enhance the effects of retrieval practice in real educational settings. Theoretical accounts assume retrieval practice to be the most efficient whenever retrieval is difficult but successful. Therefore, we developed a novel retrieval practice procedure for multiple-choice questions that adapts to learners' abilities and can be applied irrespective of learning content. This adaptive retrieval practice procedure aims to make retrieval gradually easier whenever students provide an incorrect answer. In a field experiment, students read book chapters which served as learning content as part of a weekly university course. In three consecutive weeks, they then practiced this weeks' reading assignment by (a) adaptive testing, (b) non-adaptive testing, and (c) restudy in counter-balanced order. In Week 4 a surprise criterial test took place. On average, restudy outperformed both testing conditions, whereas adaptive testing performed equally well as non-adaptive testing. However, exploratory analyses revealed that with increasing retention intervals, the superiority of restudy disappeared. Furthermore, whenever participants fully read the assigned chapters and retention intervals increased, adaptive testing outperformed non-adaptive testing. In sum, adaptive retrieval practice did not prove to be generally superior to non-adaptive retrieval practice or restudy but retention interval and students' preparation for class might be conditions rendering adaptive retrieval useful in educational settings.

Chapter V

Learners and lecturers often use computer-assisted techniques to revise learning content. Conventional techniques include the use of (electronic) flashcards and clicker questions in offline courses (Caldwell, 2007; Golding, Wasarhaley, & Fletcher, 2012; Mayer et al., 2009; Wissman, Rawson, & Pyc, 2012), or quizzes in massive open online courses (MOOC; Chauhan, 2017). Digital flashcards and online quizzes are self-directed learning procedures in which learners respond to questions about the learning content. Clicker questions are used in classroom settings and are usually provided by the instructor and immediately answered by the learners. Learners using these technologies, knowingly or unknowingly benefit from the testing effect, also known as retrieval practice effect or test-enhanced learning. The testing effect means that practicing learned content by an active retrieval from memory is more beneficial for retention than restudying the same learning content. This testing effect has been reliably found in many laboratory studies (cf. the meta-analyses by Adesope, Trevisan, & Sundararajan, 2017; Phelps, 2012; Rowland, 2014). Furthermore, empirical evidence indicates that the testing effect can be fruitfully applied to real-world educational contexts (see the meta-analyses by Adesope et al., 2017; Bangert-Drowns, Kulik, & Kulik, 1991; Schwieren, Barenberg, & Dutke, 2017).

The strong evidence for the testing effect in improving learning outcomes from laboratory studies has sparked research on how to maximize the effects, although with limited results. Despite successful demonstrations in the laboratory of how the testing effect can be increased, the practical impact of these improvements seems to be limited to specific learning content (e.g. vocabulary) or it requires complex schedules.

In the following review, we outline an approach that might, in principle, improve the benefits of the testing effect for all learning content on a single testing occasion. We first

present theoretical underpinnings of this approach before describing the study that is designed to test this approach in an existing university course.

Factors Influencing the Effectivity of the Testing Effect in Educational Settings

In their seminal study, Roediger and Karpicke (2006, Experiment 2) demonstrated that repeated testing of studied information leads to better retention than repeated restudy. They further demonstrated that these results occurred after two days and after one week. This testing effect has been repeatedly found in laboratory and applied contexts alike, and researchers consequently advise the use of tests in educational settings (Dunlosky & Rawson, 2015; Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013; Dunn, Saville, Baker, & Marek, 2013). Recent research has primarily focused on the use of testing schedules (Lindsey, Shroyer, Pashler, & Mozer, 2014; Rawson & Dunlosky, 2012; Rawson, Dunlosky, & Sciartelli, 2013) to enhance student outcomes. However, little is known about the optimal implementation of unique testing sessions that teachers and students can employ such as computer-assisted tests at the end of course sessions in online courses or in preparation for exams.

To improve the testing effect, one important factor to consider is the cognitive effort needed to retrieve learning content from long-term memory. The desirable difficulties framework (R. Bjork, 1994) postulates that testing must be sufficiently difficult, and the learner needs to invest a sufficient amount of effort to successfully retrieve the relevant information to benefit long-term retention. In support of this framework, research has shown that more effortful retrieval promotes retention (Pyc & Rawson, 2009) and that retrieval effort might be a more decisive factor for the effectiveness of testing compared to retrieval

success, that is, whether the retrieved information is correct (Kornell, Klein, & Rawson, 2015).

To this end, researchers often use test items of varying difficulty to manipulate retrieval effort experimentally, and the stimulus material is administered to complete groups of learners (Carpenter, 2009; Pyc & Rawson, 2009). However, this procedure has disadvantages because the effect of difficulty on retrieval effort depends on the individual ability of the learner. Individual ability in the context of this study refers to the accessibility of initially learned information in memory. The more accessible the information, the less effort is needed to retrieve it from memory and the more likely it is retrieved successfully. In line with many theoretical accounts of the testing effect, such as the desirable difficulties framework (R. Bjork, 1994), the new theory of disuse (R. Bjork & Bjork, 1992), or the retrieval effort hypothesis (Pyc & Rawson, 2009), accessibility to information is directly linked to advantages in retrieval. Lower accessibility to information is associated with more effort needed to retrieve the information, leading to better retention of the successfully retrieved information. In other words, learners profit the most from retrieval practice when retrieval is both effortful and successful. Both parameters are determined by antecedent factors that increase learners' retrieval ability.

Research has shown that learners' ability to retrieve studied information is influenced by prior knowledge (Schneider, Gruber, Gold, & Opwis, 1993) and the time between initial study occasion and retrieval attempt (Woźniak, Gorzelańczyk, & Murakowski, 1995). Furthermore, it can be assumed that study behavior (i.e., depth of mental processing) directly affects learners' ability to retrieve the studied information (Craik & Lockhart, 1972). Given the many factors that influence learners' ability to retrieve information, effortful and successful retrieval varies strongly in real world educational

contexts. The high variability suggests the use of an adaptive approach that tailors item difficulty to the ability level of students. Minear, Coane, Boland, Cooney, and Albat (2018) recently investigated the effects of student characteristics (fluid intelligence and vocabulary knowledge) and item difficulty on the testing effect in vocabulary learning. The strongest testing effects were observed for items that matched students' abilities. Participants with low fluid intelligence and vocabulary knowledge profited the most from retrieving easy items from memory, whereas participants with high fluid intelligence and vocabulary knowledge profited the most from difficult items. The authors interpret these effects as a result of a match between participants' abilities and the retrieval difficulty. However, it is noteworthy that in this study item difficulty was not adjusted, and thus the beneficial effects in each group of learners applied only to a subset of items. An alternative approach that bears the potential to maximize the testing effect would be to tailor every item to learners' ability.

One approach to systematically tailoring item difficulty to learners' ability level is altering the informativeness of retrieval cues in testing conditions. Previous work has shown that less informative cues led to higher retrieval difficulty and thus to more pronounced testing effects (Carpenter & DeLosh, 2006; Carroll & Nelson, 1993; Finley, Benjamin, Hays, Bjork, & Kornell, 2011). In this paradigm, cue informativeness is usually manipulated by altering the number of target-word letters when practicing retrieval of single words (e.g., in vocabulary learning). Fiechter and Benjamin (2017) report differential effects of cue informativeness for different levels of learners' abilities. At low ability levels, higher cue informativeness led to a higher testing effect. However, participants in this study received all cue levels irrespective of actual participants' ability levels. Thus, item difficulty was not adapted to participants' abilities.

Finn and Metcalfe (2010) followed a different approach. Participants were presented with short-answer trivia questions. Whenever an incorrect answer was entered, one of four types of feedback was given: (1) correct response (*standard feedback*), (2) opportunity to enter another answer (*minimal feedback*), (3) same question in an answer-until-correct multiple-choice format (*answer until correct*), or (4) opportunity to enter as many new answers as needed until the question was answered correctly. For each incorrect answer, a cue in the form of one letter of the target word appeared (*scaffolded feedback*). With these features, the scaffolded feedback condition represents an adaptation of cue informativeness to participants' ability levels. This condition outperformed all other conditions on retention of the correct answer after retention intervals of 0.5 hr and 24 hr. However, these findings cannot be readily generalized to the current research question. First, the study lacked a restudy control, which precludes the interpretation of a testing effect. Second, two possible confounds hamper the conclusion that adaptive testing is more beneficial than non-adaptive testing: (1) When comparing the scaffolded feedback condition to the standard feedback condition, the findings may be confounded with the time spent on learning. In the scaffolded feedback condition, participants were exposed to the question and cues until they provided the correct answer, whereas in the standard feedback condition, exposure ended after the correct answer had been shown; (2) When comparing scaffolded feedback to the answer-until-correct condition, the findings can be confounded by the change in question format. That is, answering multiple choice questions might lead to smaller testing effects than short-answer questions (for a review, see Karpicke, 2017). Finally, answers to the questions used in this study consisted of only one word. Students normally encounter complex learning content in such educational contexts. Thus, application of these findings to such contexts is limited.

Despite its limitations, the method used by Finn and Metcalfe (2010) provides further opportunities for exploring ways to match learners' ability to retrieval difficulty. To adapt this approach to real-world learning contexts, the main change involves the question format. Multiple-choice items allow for numerous response options, which provides the possibility of using new approaches involving the use of multimedia and response options that differ from mere descriptions of the correct answer (e.g., Davey, Godwin, & Mittelholtz, 1997; Parshall, Stewart, & Ritter, 1996). Furthermore, feedback on multiple-choice responses can be provided immediately in computer-assisted learning environments, making multiple-choice items particularly suitable for adaptive computerized learning (e.g., Martin & Lazendic, 2018; Parshall, Spray, Kalohn, & Davey, 2002).

Similar to studies that varied cue informativeness by increasing the number of target-word letters, we propose a procedure that varies cue informativeness by reducing the number of selectable response options. Both procedures are assumed to promote correct answers by increasing the probability of guessing correctly, but more importantly, current procedural accounts on the testing effect state that reducing the set of possible candidates of a cue-target connection strengthens the remaining cue-target connections (Grimaldi & Karpicke, 2012). Therefore, constraining the set of possible responses in both procedures leads to better memory for the remaining possible response options. Furthermore, incorrect options in the proposed procedure are not only deleted from the set of selectable response options but are also marked as incorrect. The latter clearly adds information, thus increasing the cue informativeness.

An ongoing debate questions whether multiple-choice items produce testing effects similar to the effects produced by short-answer questions (for a review, see Karpicke, 2017). Numerous studies have suggested that multiple-choice testing compared to short-answer

testing might lead to inferior testing effects (Kang, McDermott, & Roediger, 2007), equal testing effects (McDaniel, Wildman, & Anderson, 2012; Smith & Karpicke, 2014), or even superior testing effects (Little, Bjork, Bjork, & Angello, 2012). Karpicke (2017) discussed the possibility that different retrieval difficulties in multiple-choice and short-answer items might lead to these inconsistent findings. Consequently, matching learners' abilities and retrieval difficulty with multiple-choice items might augment testing effects.

Rationale of this Study

Previous research has shown that retrieval practice can be fruitfully applied to computer-assisted learning in educational contexts (e.g., Cook, Thompson, & Thomas, 2014; Cook, Thompson, Thomas, Thomas, & Pankratz, 2006; DelSignore, Wolbrink, Zurakowski, & Burns, 2016; Friedl et al., 2006; Grimaldi & Karpicke, 2014; Kerfoot, DeWolf, Masser, Church, & Federman, 2007; Maag, 2004; Schmidmaier et al., 2011; Shapiro & Gordon, 2012). In short, retrieval practice using multiple-choice questions can benefit learning. When the correct answers are single word, retrieval practice is most beneficial when participants' abilities match items with the optimum amount of cue informativeness. Given these preliminary findings and the theoretical accounts on the testing effect, adapting the difficulty of each item to learners' abilities might benefit retention more than standard testing procedures.

The aim of this study is to compare a procedure that adapts retrieval cue informativeness to learners' ability levels to standard procedures of retrieval practice and then examine the potential of this adaptive testing procedure for complex learning content. To this end, we developed a novel adaptive testing procedure for multiple-choice questions which allows us to investigate the beneficial effects of adaptive retrieval practice in an existing university course.

We manipulated students' practice strategies after they visited a university course session. Practice consisted of (a) testing in which cue informativeness adapted to learners' ability levels, (b) testing in which no adaptation of cue informativeness took place, or (c) restudying as a control condition. Testing included multiple-choice items, and cue informativeness was operationalized by providing feedback on incorrect response options to the learner. We assessed the effectiveness of practice strategies by means of a surprise criterial test administered between one and seven days after the last practice session. We also assessed learners' effort in practicing the course content. We expected both testing conditions to be superior to restudy (*testing effect hypothesis*) and adaptive testing to be superior to non-adaptive testing (*adaptive testing effect hypothesis*).

Method

Participants, Power, and Required Sample Size

Participants were recruited from two university courses attending a course on behavioral disorders. The students are enrolled in a teacher training program and will eventually become teachers in different school forms. To our knowledge, Fiechter and enjamin (2017) conducted the only study investigating adaptive testing compared to non-adaptive testing and restudying. They reported effect sizes (Cohen's *d*) between 0.28 (Experiments 1a–1e) and 0.51 (Experiments 2a–2b) for the difference between the two testing conditions. The experiments in this study implemented different conditions, none of which suitably match our research question. We thus used the weighted mean of these effect sizes ($M = 0.41$) as the basis for an a priori power analysis with a required power of $1\text{-}\beta = .90$. Power analysis was conducted with the tools provided by Judd, Westfall, & Kenny (2017). For a within-participants design (see the Design section), this implies a minimum of 46

participants to detect a significant difference between the two testing conditions. Regular course size in the target population ranges between 35 and 40 students. Thus, students from two courses were asked to participate in exchange for course credit. In this semester, students chose from a total of seven courses on this topic, whereas only these two courses included participation in a study to fulfill course credit. Participants gave their informed and written consent prior to participation.

A total of 68 students (72% female) took part in the study. Participants' age ranged from 18 to 31 years (M = 21.04, SD = 2.49) and participants were mostly students in their first term (M = 1.53, SD = 1.08). The procedures for analyzing the data can handle missing data, hence we did not exclude data from participants with partially missing data. Whenever participants failed to show up for their practice sessions or technical errors occurred that lead to data loss during the experiment, we used the remaining data points. We assumed that any missing data points will be missing completely at random and thus inferences can proceed by analyzing only the observed data (Ibrahim & Molenberghs, 2009).

Procedure

General procedure

The study was conducted in the last weeks of the semester. Participants were advised to read book chapters in preparation for the course sessions. All course sessions were taught by the first author, and course content was largely based on the reading assignments. Three subsequent course sessions addressed the topics and practice sessions were offered, which were subject to manipulation (i.e., the focal sessions). After each focal session, participants were asked to practice the course content of the last session in the laboratory within one week. Participants returned to the laboratory within one week after the session that follows the last focal session, ostensibly to practice one additional session but instead the surprise criterial test was administered.

Practice sessions and criterial test

In each practice session, participants first answered sociodemographic items, questions about their presence in the course session, questions about prior knowledge in the domain of the focal session, and questions concerning whether and when the reading assignment was completed. Participants then engaged in practicing the course content according to one of the three practice conditions (adaptive testing, non-adaptive testing, or restudy). Practice was self-paced and consisted of five rounds. In each round, all information units were practiced in randomized order.

In the *restudy condition,* statements were the same in each round. In both testing conditions*,* each round consisted of fill-in-the-blank items with two blanks (see section Materials). In the *adaptive testing condition*, the items were the same in each round. However, participants' performance on each item affected the difficulty of this question in subsequent rounds. Every time an item was answered incorrectly, one response option was permanently eliminated from the question. Response options from both blank spaces were eliminated alternatingly. Each elimination decreased the amount of possible incorrect combinations of response options. The resulting combinations for Rounds 1–5 when all responses were incorrect were 15 (without elimination), 11, 8, 5, and 3, respectively. Eliminated options were still visible but could not be selected. Whenever a response option has been eliminated, in subsequent rounds a note appeared on the screen reminding the participants to reflect why the eliminated options might be wrong and then consider their self-generated reasons when attempting to retrieve the correct option. In the *non-adaptive testing condition*, the items were identical in each round and the amount of selectable and eliminated response options were each set to two. Instead of being adaptive, the practice test thus always provided the maximal level of cue informativeness.

After each test, participants were asked to rate the difficulty of the item on a visual analogous scale, ranging from “very easy” to “very difficult”. In all conditions following each information unit in each round, participants were asked to predict retention of the information unit on a visual analogous scale, ranging from “very good” to “very bad.”

In both testing conditions, a message then indicated whether an item was answered correctly.

At the beginning of the criterial test, participants were informed that no further practice would take place and that they would be tested on the three previous course sessions. All items were then presented in randomized order and without a time limit. Finally, participants were thanked, debriefed and reminded not to disclose information regarding this study to other students.

Materials

Test items and restudy statements

Three chapters from a textbook on mental disorders that are part of the regular reading assignments of the course were selected as the basis for study material. The content of the chapters on “Drug abuse and addiction,” “Suicidality,” and “Affective Disorders” were surveyed, and 30 information units per topic were identified. For each information unit, one statement and one fill-in-the-blank item were created by summarizing the key information of the information unit. An example statement is: “Massive intoxications lead to absence of positive states of mind (“highs”). With longer duration of the addiction, the proportion of positive effects on the mind of the user decreases whereas the proportion of poisonous outcomes increases.” Fill-in-the-blank items were created by asking for the key information from the information unit by leaving two blank spaces and providing four response options for each blank space, for example:

"Massive intoxications lead to absence of _____(Blank 1)_____. With longer duration of the addiction, the proportion of positive effects on the mind of the user decreases, while _____(Blank 2)_____increases."

Options for Blank 1: (A) positive states of mind ("highs"), (B) cravings for the substance, (C) resistance of the blood-brain barrier, (D) refractory periods of involved neurons.

Options for Blank 2: (A) the proportion of dysphoric intrusions, (B) the proportion of poisonous outcomes, (C) the desire for abstinence, (D) the proportion of abstinent periods.

Answers were scored as correct answers only when the correct response options for both blank spaces were selected, which corresponds to one combination of response options out of 16.

Practice materials

For each practice session, 20 information units were randomly drawn from the 30 information units prepared for this session. Based on the selected information units, materials were prepared for each practice session. The materials were presented with the software Inquisit 5 (Version 5.0.6.0; Millisecond Software, 2016) and consisted of either 20 fill-in-the-blank items (adaptive testing and non-adaptive testing condition) or 20 summarizing statements (restudying). In all three conditions, each information unit was presented on one page. The presentation order of fill-in-the-blank items was randomized on each practice session.

Criterial test

A criterial test was constructed that consisted of 20 items from each topic. For each topic, 10 items were based on information units used in the practice material, and 10 items

were based on information units not used in the practice material. Each criterial test item was presented along with four response options with only one correct answer.

Design

We investigated the effect of the independent variable practice condition (adaptive testing, non-adaptive testing, restudy) across three course sessions on the dependent variable performance in the criterial test. All participants experienced all practice conditions in the course sessions (within-participants design). To prevent effects of topic and sequence, we counterbalanced the sequence of conditions, thus resulting in a total of six combinations of conditions and topics. Table V.1 illustrates the possible combinations of conditions across the topics. Each participant was randomly assigned to one of these six combinations upon arrival at the first practice session.

Table V.3

Possible Combinations of Practice Conditions

	Topic		
Combination Nr.	Suicidality	Drug Abuse and Addiction	Affective Disorders
1	Restudy	Adaptive testing	Non-adaptive testing
2	Restudy	Non-adaptive testing	Adaptive testing
3	Adaptive testing	Non-adaptive testing	Restudy
4	Adaptive testing	Restudy	Non-adaptive testing
5	Non-adaptive testing	Restudy	Adaptive testing
6	Non-adaptive testing	Adaptive testing	Restudy

Note. Combinations of practice conditions across course sessions (sequence of topics/conditions were counterbalanced across participants)

Results

We estimated generalized linear mixed effect models (GLMMs) with a logit-link function (Dixon, 2008) and linear mixed effect models with the R package lme4 (Bates, Mächler, Bolker, & Walker, 2015). Mixed effect models have many advantages compared to ANOVAs (e.g., see Baayen, Davidson, & Bates, 2008; Richter, 2006). These advantages include better options for analyzing categorical outcome variables (Jaeger, 2008) and for dealing with missing data. The package emmeans (formerly: lsmeans) was used (Lenth, 2016) for comparisons between experimental conditions and estimating performance scores for different conditions. Type I error probability was set to .05 for all significance tests. The multivariate *t* distriConfirmatory Analyses Regarding the Testing Effect Hypothesis and the Adaptive Testing Effect Hypothesisbution was used to adjust *p* values (for details, see Lenth, 2016) for post-hoc tests (but not for planned comparisons). Participants and test items were included as random effects (random intercepts) in all models.

Criterial tests were scored with 1 when the correct option was ticked vs. 0 when a distractor was ticked. All models were estimated on the item level (items x participants) of either the criterial test or the practice material.

Confirmatory Analyses Regarding the Testing Effect Hypothesis and the Adaptive Testing Effect Hypothesis

We used Helmert coding to create two orthogonal contrasts that correspond to the hypotheses: The first contrast compared the two testing conditions (coded with -1) to the restudy condition (coded with 2) and thus evaluated the testing effect hypothesis. The second contrast compared the adaptive testing condition (coded with 1) to the non-adaptive testing condition (coded with -1) and thus evaluated the adaptive testing effect hypothesis; the restudy condition was coded with 0 in this latter contrast. We estimated a model including both contrasts as predictors and the probability of providing a correct response in the criterial

test as dependent variable. The model estimates are shown in Table V.2. Results revealed a negative effect of testing compared to restudying and no difference among the testing conditions. Overall, adaptive testing (P = .42, SE = .04, z = -3.51, p = .001, OR = 0.74) as well as non-adaptive testing (P = .43, SE = .04, z = -2.92, p = .010, OR = 0.78) lead to lower probabilities of answering correctly than restudying (P = .49, SE = .04) The estimated probabilities in the testing conditions did not differ significantly from each other (z = 0.583, p = .415, one-tailed, OR = 1.05).

Table V.4

Model Parameters for Hypothesis Testing

Parameter	β	*SE*	*z*	*p*
Intercept	-0.22	0.14	-1.59	.112
Testing vs. Restudy	0.10	0.02	3.72	< .001
Adaptive Testing vs. Non-Adaptive Testing	-0.03	0.04	-0.58	.600
$N_{\text{Participants}}$		68		
N_{Items}		60		

Note. Parameter estimates for the models estimating the effect of testing and the effect of adaptive testing realized by two orthogonally coded contrasts (helmert coding). Testing vs. restudy (contrast-coded: adaptive testing = -1, non-adaptive testing = -1, restudy = 2). Adaptive testing vs. non-adaptive testing (contrast-coded: adaptive testing = 1, non-adaptive testing = -1, restudy = 0).

Exploratory Analyses

For further exploratory analyses, investigating potential moderators of the testing effect and the adaptive testing effect we considered a set of exploratory predictors that might arguably be involved in both effects. We expected an interplay of participants' abilities and benefits of practice procedures and expected participants' abilities to be a result of the study behavior. Specifically, as most theoretical accounts on the testing effect state, abilities should affect the testing effect by altering the difficulty of retrieval (e.g., Carpenter, 2009; Pyc & Rawson, 2009). As one moderator, we considered self-reported fulfillment of reading assignments with the three levels "no reading", "partial reading", and "full reading" of the assigned chapters (Helmert-coded). For the same reason, we considered self-reported presence in the course session with the two levels "present" and "absent" (dummy-coded: absent = 0, present = 1) as a second predictor. Theoretical accounts on the testing effect often assume more difficult practice procedures to result in more sustainable memory traces (e.g.,

Roediger & Karpicke, 2006a, 2006b; Rowland, 2014). Therefore, the retention interval, that is the time interval between the lab session and the criterial test centered around the mean (*M* = 17.73) was included in days. All these predictors were included as participant-level predictors and could vary for each topic. We estimated separate models for differences between testing and restudying (contrast-coded: testing conditions = -1, restudy condition = 2) and for differences between the testing conditions (dummy-coded: adaptive testing = 1, non-adaptive testing = 0). We estimated multiple models using the probability of answering correctly as dependent variable and included different combinations of this set of predictors. However, for each effect we will only present the most parsimonious model that includes only the significant interaction effects. Due to the exploratory nature of these analyses, all moderator effects were tested with two-tailed tests.

Moderators of the testing effect

The most parsimonious model involving moderators of the testing effect revealed a negative effect of testing compared to restudying and a positive effect of the retention interval on performance in the criterial test (Table V.3). More importantly, there was a significant interaction between the learning condition and the retention interval: The longer the retention interval, the more beneficial became testing compared to restudying. Figure V.1 depicts this interaction. Post-hoc comparisons revealed that restudying outperformed testing in the whole range from the minimum retention interval of one day (ΔP = -.19, *SE* = .06, *z* = -3.22, *p* = .001, *OR* = 0.43) to a retention interval of 20 days (ΔP = -.05, *SE* = .02, *z* = -2.42, *p* = .016, *OR* = 0.82). However, this difference became insignificant with longer retention intervals from 21 days (ΔP = -0.04, *SE* = 0.02, *z* = -1.86, *p* = .063, *OR* = 0.85) to the maximum retention interval of 29 days (ΔP = .03, *SE* = .05, *z* = .74, *p* = .458, *OR* = 1.11).

Table V.5

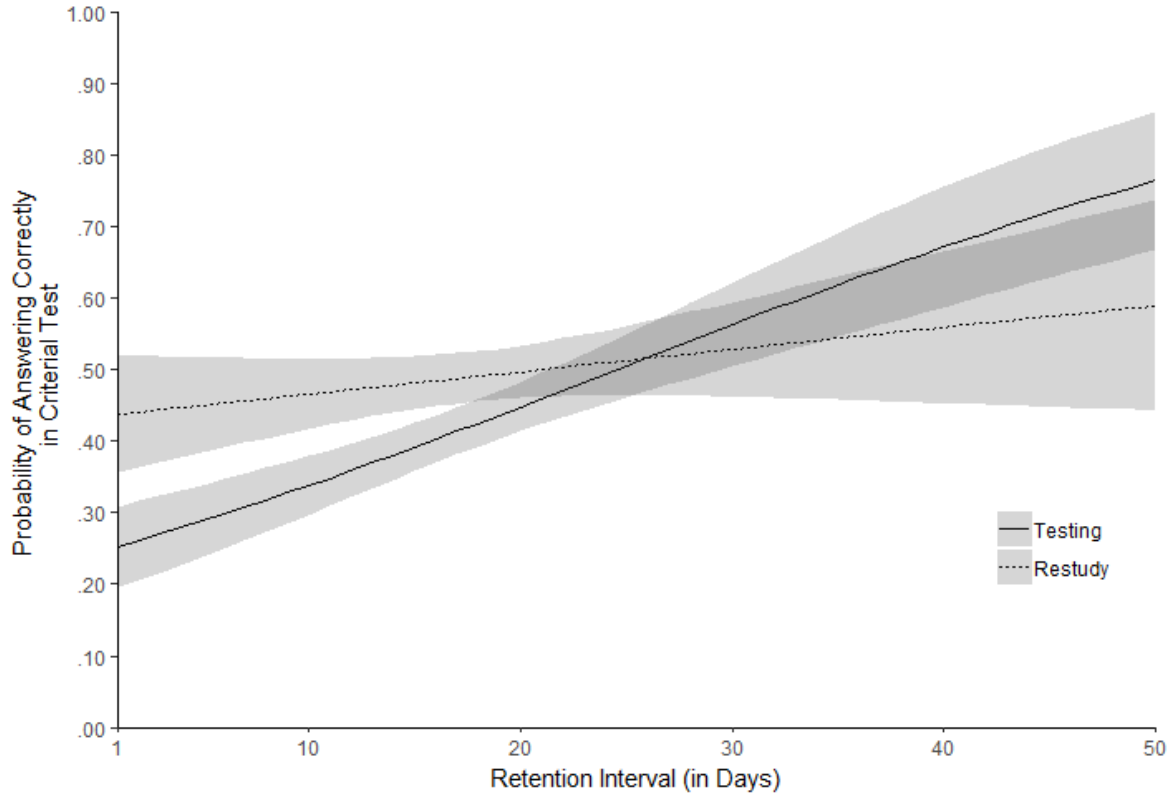
Model Parameters for the Exploratory Testing Effect Model

Parameter	β	*SE*	*z*	*p*
Intercept	-0.31	0.13	-2.35	.019
Practice Condition	0.27	0.07	3.68	<.001
Retention Interval	0.05	0.02	2.88	.004
Practice Condition x Retention Interval	-0.03	0.01	-2.42	.015
$N_{\text{Participants}}$		68		
N_{Items}		60		

Note. Parameter estimates for the most parsimonious model including moderators of the testing effect. Practice condition (contrast-coded: adaptive testing = -1, non-adaptive testing = -1, restudy = 2). Retention interval (centered around $M = 17.73$).

Figure V.1.

The influence of retention interval on the testing effect



Note. Probability of correct responses in criterial test items (back-transformed from the logits in the GLMM) by retention interval and testing condition (adaptive testing vs. non-adaptive testing). Areas around the graphs indicate standard errors.

Moderators of the adaptive testing effect

The most parsimonious model involving moderators of the adaptive testing effect included the full set of exploratory predictors (Table V.4). We observed no main effect of the testing condition on criterial test performance. Testing conditions interacted positively with the retention interval and negatively with the presence in the course session. This indicates that adaptive retrieval practice was more beneficial for longer retention intervals and that non-adaptive retrieval practice was more beneficial when participants visited course sessions prior to being tested. Furthermore, there was a three-way interaction of the testing condition with retention interval and fulfillment of the reading assignment: Whenever participants fully

read the assigned chapters and retention interval increased, adaptive testing was more beneficial. Most notably, post-hoc comparisons revealed significant differences between adaptive and non-adaptive testing in the probability of providing a correct response in the criterial tests for participants who fully read the assigned chapters: At the maximum retention interval of 29 days and more, adaptive testing outperformed non-adaptive testing, irrespective of participants being present ($\Delta P = 0.28$, $SE = 0.12$, $z = 2.35$, $p = .019$, $OR = 2.55$) or absent in the course session ($\Delta P = 0.56$, $SE = 0.18$, $z = 3.09$, $p = .002$, $OR = 15.00$).

Table V.6

Model Parameters for the Exploatory Adaptive Testing Effect Model

Parameter	β	*SE*	*z*	*p*
Intercept	-0.96	0.27	-3.53	<.001
Testing Condition	0.56	0.31	1.80	.072
Partly Reading vs. No Reading	-0.01	0.13	-0.07	.945
Full Reading vs. Reading Less	-0.38	0.23	-1.63	.103
Retention Interval	0.01	0.02	0.33	.739
Presence in Course Session	0.97	0.29	3.33	<.001
Testing Condition x Retention Interval	0.05	0.02	2.84	.005
Testing Condition x Partly Reading vs. No Reading	0.15	0.16	0.93	.355
Testing Condition x Full Reading vs. Reading Less	0.38	0.30	1.26	.209
Testing Condition x Presence in Course Session	-0.86	0.36	-2.39	.017
Partly Reading vs. No Reading x Presence in Course Session	0.07	0.17	0.44	.658
Full Reading vs. Reading Less x Presence in Course Session	0.55	0.28	1.98	.047
Partly Reading vs. No Reading x Retention Interval	-0.01	0.01	-0.54	.587
Full Reading vs. Reading Less x Retention Interval	-0.03	0.01	-2.20	.028
Testing Condition x Partly Reading vs. No Reading x Retention Interval	0.01	0.02	0.74	.460
Testing Condition x Full Reading vs. Reading Less x Retention Interval	0.04	0.02	2.24	.025
Testing Condition x Partly Reading vs. No Reading x Presence in Course Session	0.03	0.22	0.12	.906
Testing Condition x Full Reading vs. Reading Less x Presence in Course Session	-0.46	0.34	-1.33	.183
$N_{Participants}$		68		
N_{Items}		60		

Note. Parameter estimates for the most parsimonious model including moderators of the adaptive testing effect. Testing condition (dummy-coded: adaptive testing = 1, non-adaptive testing = 0). Retention interval (centered around M = 17.73). Partly reading vs. no reading (contrast-coded: “Read parts” = 1, “Read nothing” = -1, “Read everything” = 0). Full reading vs. reading less (contrast-coded: “Read parts” = -1, “Read nothing” = -1, “Read everything” = 2). Presence in course session (dummy-coded = “Present” = 1, “Not present” = 0).

Discussion

We designed a novel procedure for practicing adaptive retrieval to increase the benefits of the testing effect in a university course. In this procedure, retrieval was gradually made easier until participants answered the question correctly. The adaptive retrieval practice procedure was based on theoretical accounts of the testing effect that state that in order to be most effective, retrieval needs to be both successful and sufficiently difficult (Pyc & Rawson, 2009; R. Bjork, 1994; R. Bjork & Bjork, 1992).

We compared adaptive retrieval practice to restudy and to a non-adaptive practice procedure, in which all questions were always presented in the easiest form. We expected both testing conditions to be superior to restudying (testing effect hypothesis) and adaptive testing to be superior to non-adaptive testing (adaptive testing effect hypothesis). Contrary to our assumptions, restudying overall led to better retention than retrieval practice and no differences between the testing conditions were observed.

In subsequent exploratory analyses, we investigated the role of potential moderators on the testing effect and the adaptive testing effect. For the testing effect, the retention interval moderated the differences between retrieval practice and restudying: Results indicated that with longer retention intervals the benefits of retrieval practice on retention increased, while the benefits from restudying decreased. This finding is in line with many studies investigating the role of the retention interval on the testing effect (e.g., Roediger & Karpicke, 2006a, 2006b; Rowland, 2014; Toppino & Cohen, 2009; Wheeler, Ewers, & Buonanno, 2003). Furthermore, it has been shown that higher proportions of unretrievable items in retrieval practice lead to higher benefits of restudying in the short run (Jang, Wixted, Pecher, Zeelenberg, & Huber, 2012). This finding is also in line with the bifurcation model (Kornell, Bjork, & Garcia, 2011), which postulates restudy as being more beneficial than

retrieval practice whenever retrieval success is below 50% (Rowland, 2014; for supportive evidence from a field experiment conducted in a university course, see Greving & Richter, 2018). It is thus possible that the pattern of results obtained for the testing effect in the present study was obtained because the retrieval practice procedures consisted of many items that were not successfully retrieved.

For the adaptive testing effect, the exploratory analyses revealed three moderators: Presence in the course session, self-reported fulfillment of the reading assignment, and retention interval.

Contrary to what one might expect, presence in the course session increased the beneficial effects of non-adaptive retrieval practice as compared to adaptive retrieval practice, irrespective of fulfillment of reading assignment. In this context, it is important to note that the course sessions taught and summarized the main concepts that were also included in the reading assignments. As discussed before, retrieval success was low, which might indicate that participants' abilities were low in general. Presence in the course session might have lifted participants' abilities to a level sufficient to capitalize on the benefits of the non-adaptive testing condition, which was the easiest testing condition and therefore matched participants' ability level.

Controlling for the adverse effects of presence in the course session revealed two other moderators that increased benefits of adaptive testing: Only if participants read the entire book chapter that was subject to studying, adaptive retrieval practice was superior to non-adaptive retrieval practice. We assumed that whenever test difficulty matches learners' abilities, the testing effect is the strongest. In terms of cue informativeness, adaptive testing included the most difficult questions, whereas non-adaptive testing consisted of the easiest questions only. In order to match the comparably more difficult questions in the adaptive

testing conditions, participants' ability levels needed to be high. We argue that fulfillment of the reading assignment leads to higher levels of ability which might explain the observation that beneficial effects of adaptive testing arose only if reading assignments were fulfilled. This finding is consistent with our assumptions about the benefits of the match between question difficulty and learners' abilities. Furthermore, the most positive effects of adaptive retrieval practice as compared to non-adaptive retrieval practice were obtained when retention intervals increased.

Recent research from other labs has shown adaptive retrieval practice to benefit learners in terms of efficient diagnosis of students' abilities and motivation to take tests (Martin & Lazendic, 2018; Morphew, Mestre, Kang, Chang, & Fabry, 2018). In a study investigating the benefits of adaptive retrieval practice compared to non-adaptive retrieval practice, adaptive retrieval practice produced higher testing effects than non-adaptive retrieval practice (Heitmann, Grund, Berthold, Fries, & Roelle, 2018). In this study, participants first saw an e-lecture before answering easy (Level 1, reproduction of singular information unit) to difficult (Level 4, application of multiple information units) questions about the contents of the e-lecture. The sequence of these questions was either fixed (non-adaptive testing) or depended on the correctness of participants' responses, which in turn was rated by the participants themselves. The authors furthermore reported that the beneficial effects of adaptive testing depended on the performance in testing, which can be seen as a measure of students' ability. In sum, the findings from this study provides further evidence for the assumption that adaptive retrieval practice can be fruitfully applied to improve the benefits of retrieval practice, whenever students differ in their abilities.

Along the same line of reasoning, the lack of general benefits of adaptive testing over non-adaptive testing and the superiority of the restudy condition might be attributed to

the overall low level of students' abilities. Future research should follow up on this issue by investigating adaptive retrieval practice in student samples with a broader range of abilities, including higher levels of ability. Another limitation that the study shares with other field experiments concerns potential external influences (e.g., metacognitive or motivational factors, students' learning activities outside the lab) that potentially play a much greater role for performance in the criterial tests than in typical laboratory experiments on retrieval practice.

We demonstrated in this study that in some cases an adaptive retrieval practice procedure was more beneficial than non-adaptive retrieval practice. In regard to the practical implications it should be noted, that this procedure was implemented in an existing university course. Whenever students prepared for the course, they benefitted from adaptive testing more than from non-adaptive testing and the benefits increased in the long run. In real-world educational settings, practitioners have limited influence on the abilities of students prior to practicing retrieval. However, retention intervals in such settings are usually long. Thus, instructors should support their students to prepare for the course and combine these efforts with adaptive tests in an attempt to increase the retention over longer periods of time.

To conclude, in this research we developed a novel, scalable adaptive retrieval practice procedure for multiple-choice questions which failed to show its general effectiveness as compared to non-adaptive testing and restudy. However, we identified potential moderators and conditions that made this adaptive retrieval practice procedure beneficial. In this regard, this study contributes to advancing the research of increasing the benefits of retrieval practice procedures.

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*, 659–701. https://doi.org/10.3102/0034654316689306

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390–412. https://doi.org/10.1016/j.jml.2007.12.005

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

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67.* https://doi.org/10.18637/jss.v067.i01

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). Cambridge, MA: MIT Press.

Caldwell, J. E. (2007). Clickers in the large classroom: Current research and best-practice tips. *CBE—Life Sciences Education*, *6*, 9–20. https://doi.org/10.1187/cbe.06-12-0205

Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: The benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*, 1563–1569. https://doi.org/10.1037/a0017021

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

Carroll, M., & Nelson, T. O. (1993). Failure to obtain a generation effect during naturalistic learning. *Memory & Cognition*, *21*, 361–366. https://doi.org/10.3758/BF03208268

Chauhan, J. (2017). Quiz in MOOC: An overview. *International Research Journal of Engineering and Technology (IRJET)*, *4*, 303–307.

Cook, D. A., Thompson, W. G., & Thomas, K. G. (2014). Test-enhanced web-based learning: Optimizing the number of questions (a randomized crossover trial). *Academic Medicine, 89*, 169–175. https://doi.org/10.1097/ACM.0000000000000084

Cook, D. A., Thompson, W. G., Thomas, K. G., Thomas, M. R., & Pankratz, V. S. (2006). Impact of self-assessment questions and learning styles in web-based learning: A randomized, controlled, crossover trial. *Academic Medicine, 81*, 231–238. https://doi.org/10.1097/00001888-200603000-00005

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, *11*, 671–684. https://doi.org/10.1016/S0022-5371(72)80001-X

Davey, T., Godwin, J., & Mittelholtz, D. (1997). Developing and scoring an innovative computerized writing assessment. *Journal of Educational Measurement, 34*, 21–41. https://doi.org/10.1111/j.1745-3984.1997.tb00505.x

DelSignore, L. A., Wolbrink, T. A., Zurakowski, D., & Burns, J. P. (2016). Test-enhanced e-learning strategies in postgraduate medical education: A randomized cohort study. *Journal of Medical Internet Research, 18*, 146–154. http://dx.doi.org/10.2196/jmir.6199

Dunlosky, J., & Rawson, K. A. (2015). Practice tests, spaced practice, and successive relearning: Tips for classroom use and for guiding students' learning. *Scholarship of Teaching and Learning in Psychology*, *1*, 72–78. https://doi.org/10.1037/stl0000024

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

Dunn, D. S., Saville, B. K., Baker, S. C., & Marek, P. (2013). Evidence-based teaching: Tools and techniques that promote learning in the psychology classroom. *Australian Journal of Psychology*, *65*, 5–13. https://doi.org/10.1111/ajpy.12004

Fiechter, J. L., & Benjamin, A. S. (2017). Diminishing-cues retrieval practice: A memory-enhancing technique that works when regular testing doesn't. *Psychonomic Bulletin & Review*. https://doi.org/10.3758/s13423-017-1366-9

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

Finn, B., & Metcalfe, J. (2010). Scaffolding feedback to maximize long-term error correction. *Memory & Cognition*, *38*, 951–961. https://doi.org/10.3758/MC.38.7.951

Friedl, R., Höppler, H., Ecard, K., Scholz, W., Hannekum, A., Oechsner, W., & Stracke, S. (2006). Comparative evaluation of multimedia driven, interactive, and case-based teaching in heart surgery. *The Annals of Thoracic Surgery, 82*, 1790–1795. https://doi.org/10.1016/j.athoracsur.2006.05.118

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

Greving, S., & Richter, T. (2018). Examining the testing effect in university teaching: Retrievability and question format matter. *Frontiers in Psychology, 9*:2412. https://doi.org/10.3389/fpsyg.2018.02412

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

Grimaldi, P. J., & Karpicke, J. D. (2014). Guided retrieval practice of educational materials using automated scoring. *Journal of Educational Psychology, 106*, 58–68. https://doi.org/10.1037/a0033208

Heitmann, S., Grund, A., Berthold, K., Fries, S., & Roelle, J. (2018). Testing is more desirable when it is adaptive and still desirable when compared to note-taking. *Frontiers in Psychology, 9*. https://doi.org/10/gfrgk5

Ibrahim, J. G., & Molenberghs, G. (2009). Missing data methods in longitudinal studies: A review. *Test (Madrid, Spain)*, *18*, 1–43. https://doi.org/10.1007/s11749-009-0138-x

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*, 434–446. https://doi.org/10.1016/j.jml.2007.11.007

Jang, Y., Wixted, J. T., Pecher, D., Zeelenberg, R., & Huber, D. E. (2012). Decomposing the interaction between retention interval and study/test practice: The role of retrievability. *Quarterly Journal of Experimental Psychology*, *65*, 962–975. https://doi.org/10.1080/17470218.2011.638079

Judd, C. M., Westfall, J., & Kenny, D. A. (2017). Experiments with more than one random factor: Designs, analytic models, and statistical power. *Annual Review of Psychology, 68*, 601–625. https://doi.org/10.1146/annurev-psych-122414-033702

Kang, S. H. K., McDermott, K. B., & Roediger, H. L. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology*, *19*, 528–558. https://doi.org/10.1080/09541440601056620

Karpicke, J. D. (2017). Retrieval-based learning: A decade of progress. In John H. Byrne (Series Ed.), *Learning and Memory: A Comprehensive Reference* (2nd ed.)*: Vol 2.: Cognitive psychology of memory* (J. T. Wixted, Ed., pp. 487–514). Oxford: Academic Press. https://doi.org/10.1016/B978-0-12-809324-5.21055-9

Kerfoot, B. P., DeWolf, W. C., Masser, B. A., Church, P. A., & Federman, D. D. (2007). Spaced education improves the retention of clinical knowledge by medical students: A randomised controlled trial. *Medical Education, 41*, 23–31. https://doi.org/10.1111/j.1365-2929.2006.02644.x

Kornell, N., Bjork, R. A., & Garcia, M. A. (2011). Why tests appear to prevent forgetting: A distribution-based bifurcation model. *Journal of Memory and Language, 65*, 85–97. https://doi.org/10.1016/j.jml.2011.04.002

Kornell, N., Klein, P. J., & Rawson, K. A. (2015). Retrieval attempts enhance learning, but retrieval success (versus failure) does not matter. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*, 283–294. http://dx.doi.org/10.1037/a0037850

Lenth, R. V. (2016). Least-squares means: The R package lsmeans. *Journal of Statistical Software*, *69*. https://doi.org/10.18637/jss.v069.i01

Lindsey, R. V., Shroyer, J. D., Pashler, H., & Mozer, M. C. (2014). Improving students' long-term knowledge retention through personalized review. *Psychological Science*, *25*, 639–647. https://doi.org/10.1177/0956797613504302

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*, 1337–1344. https://doi.org/10.1177/0956797612443370

Maag, M. (2004). The effectiveness of an interactive multimedia learning tool on nursing students' math knowledge and self-efficacy. *CIN: Computers, Informatics, Nursing, 22*, 26–33.

Martin, A. J., & Lazendic, G. (2018). Computer-adaptive testing: Implications for students' achievement, motivation, engagement, and subjective test experience. *Journal of Educational Psychology, 110*, 27–45. http://dx.doi.org/10.1037/edu0000205

Mayer, R. E., Stull, A., DeLeeuw, K., Almeroth, K., Bimber, B., Chun, D., … Zhang, H. (2009). Clickers in college classrooms: Fostering learning with questioning methods in large lecture classes. *Contemporary Educational Psychology*, *34*, 51–57. https://doi.org/10.1016/j.cedpsych.2008.04.002

McDaniel, M. A., Wildman, K. M., & Anderson, J. L. (2012). Using quizzes to enhance summative-assessment performance in a web-based class: An experimental study. *Journal of Applied Research in Memory and Cognition, 1*, 18–26. https://doi.org/10.1016/j.jarmac.2011.10.001

Millisecond Software (2016). Inquisit 5 (Version 5.0.6.0) [Computer Software]. Retrieved from https://www.millisecond.com.

Minear, M., Coane, J. H., Boland, S. C., Cooney, L. H., & Albat, M. (2018). The benefits of retrieval practice depend on item difficulty and intelligence. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. http://dx.doi.org/10.1037/xlm0000486

Morphew, J. W., Mestre, J. P., Kang, H.-A., Chang, H.-H., & Fabry, G. (2018). Using computer adaptive testing to assess physics proficiency and improve exam performance in an introductory physics course. *Physical Review Physics Education Research, 14*. https://doi.org/10/gd8dqm

Parshall, C. G., Spray, J. A., Kalohn, J., & Davey, T. (2002). *Practical considerations in computer-based testing.* New York, NY: Springer Science & Business Media.

Parshall, C. G., Stewart, R., & Ritter, J. (1996). *Innovations: graphics, sound, and alternative response modes.* Presented at the Annual Meeting of the American Educational Research Association, New York, NY.

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

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

Rawson, K. A., & Dunlosky, J. (2012). When is practice testing most effective for improving the durability and efficiency of student learning? *Educational Psychology Review*, *24*, 419–435. https://doi.org/10.1007/s10648-012-9203-1

Rawson, K. A., Dunlosky, J., & Sciartelli, S. M. (2013). The power of successive relearning: Improving performance on course exams and long-term retention. *Educational Psychology Review*, *25*, 523–548. https://doi.org/10.1007/s10648-013-9240-4

Richter, T. (2006). What is wrong with ANOVA and multiple regression? Analyzing sentence reading times with hierarchical linear models. *Discourse Processes*, *41*, 221–250. https://doi.org/10.1207/s15326950dp4103_1

Roediger, H. L., & Karpicke, J. D. (2006a). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*, 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*, 181–210. https://doi.org/10.1111/j.1745-6916.2006.00012.x

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin*, *140*, 1432–1463. https://doi.org/10.1037/a0037559

Schmidmaier, R., Ebersbach, R., Schiller, M., Hege, I., Holzer, M., & Fischer, M. R. (2011). Using electronic flashcards to promote learning in medical students: Retesting versus

restudying. *Medical Education, 45*, 1101–1110. https://doi.org/10.1111/j.1365-2923.2011.04043.x

Schneider, W., Gruber, H., Gold, A., & Opwis, K. (1993). Chess expertise and memory for chess positions in children and adults. *Journal of Experimental Child Psychology, 56*, 328–349. https://doi.org/10.1006/jecp.1993.1038

Schwieren, J., Barenberg, J., & Dutke, S. (2017). The testing effect in the psychology classroom: A meta-analytic perspective. *Psychology Learning & Teaching*, *16*, 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*, 635–643. https://doi.org/10.1002/acp.2843

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

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

Wheeler, M., Ewers, M., & Buonanno, J. (2003). Different rates of forgetting following study versus test trials. *Memory, 11*, 571–580. https://doi.org/10.1080/09658210244000414

Wissman, K. T., Rawson, K. A., & Pyc, M. A. (2012). How and when do students use flashcards? *Memory, 20*, 568–579. https://doi.org/10.1080/09658211.2012.687052

Woźniak, P. A., Gorzelańczyk, E. J., & Murakowski, J. A. (1995). Two components of long-term memory. *Acta Neurobiologiae Experimentalis*, *55*, 301–305.

Chapter VI

General Discussion

The aim of this dissertation was to investigate direct testing effects in real-world educational settings and to advance the understanding of difficulty and retrievability as moderating factors. In Chapter I relevant research was reviewed and discussed and it became apparent that research investigating direct testing effects in real-world educational contexts is scarce. Also in Chapter I, I presented theoretical accounts on the testing effect that make assumptions about the complex role of test question difficulty and thus also about different question formats. For one, the bifurcation model assumes that a certain amount of information needs to be retrieved from memory in order to profit from practice tests and thus difficulty should be sufficiently low to achieve high retrievability rates (Kornell et al., 2011). This assumed link between difficulty and the extent of the testing effect seems to be at odds with the retrieval effort hypothesis (Pyc & Rawson, 2009) because the latter assumes that a higher difficulty should result in practice tests being more effective for retention. Furthermore, I argued that learners' variables interact with practice test difficulty and hence it is not only the objective difficulty of a practice test that determines its effectiveness in terms of testing effects but the match between learners' knowledge about the learned content and test difficulty.

In the following I will first summarize and discuss the key findings of the four empirical studies before discussing the theoretical and practical implications that can be derived from all studies.

Summary and Discussion of Study 1

The first empirical study presented in Chapter II investigated the possibility to elicit direct testing effects in an existing university lecture by means of a short practice test at the end of each lecture session. In this study, we compared the effects of different practice activities—namely answering multiple-choice questions, answering short-answer questions,

or restudying—on retention of lecture content 1, 12, and 23 weeks after the last practice activity took place. This study also investigated the moderating role of practice test difficulty in the form of retrievability—operationalized through is the extent an item has been correctly recalled in the practice tests.

The main findings were the superiority of short-answer testing over restudying, whenever retrievability was high and that this superiority was independent of the time retention of the lecture content to when it was assessed. Although similar designs have been employed to investigate direct testing effects for different question formats in simulated classrooms (e.g., Nungester & Duchastel, 1982) this study is the—to my knowledge—only study to report direct testing effects in real-world educational contexts. Most importantly, this effect persevered not only until the end of the semester in which practice tests were used but also until the end of the following semester, indicating a practically relevant, long-term benefit elicited by practice tests. With regard to the important role of test question difficulty, two additional findings are of importance: This study did not find evidence for a direct testing effect for short-answer questions that were retrieved by less than 46% of the participants or for direct testing effects elicited by multiple-choice questions. This finding of short-answer practice being effective only if items can be recalled correctly from memory is in line with the assumptions made by the bifurcation model, therefore in order to be effective practice tests should be easy enough to encourage successful retrieval.

It has been theorized, that answering multiple-choice questions is inapt to elicit direct testing effects because answering these questions requires less effort as only parts of the initially learned information needs to be retrieved (Glover, 1989). Findings from this first study indicate that answering multiple choice questions was indeed easy since two-thirds of multiple-choice questions—as compared to one-third of short-answer questions—were

answered correctly in most of the cases. However, there was also no beneficial testing effect for items that were harder to retrieve, indicating that some other factors (e.g., retrieval-induced forgetting, discussed in Chapter II) might render multiple-choice questions ineffective as practice activities following university lectures.

Summary and Discussion of Study 2

The second empirical study presented in Chapter III reinvestigated the research questions of the first study and further explored the possibility of rendering multiple-choice question beneficial even when no feedback was provided. This study conceptually replicated the first study while extending the research question with regard to understand the effects multiple-choice question difficulty has on retention of learning content. It has already been argued that practicing multiple-choice questions fails to elicit direct testing effects because when learners succeed in producing a correct answer, they do so by recognizing the correct response option, which does not require a complete recall of relevant information from memory (e.g., Glover, 1989). Research has demonstrated that multiple-choice questions can elicit direct testing effects whenever response alternatives are constructed in a way that requires a processing of all response options (Little & Bjork, 2015). In the second study, a similar approach was adopted in order to enhance the benefits of multiple-choice testing in an existing university lecture.

The main finding was that short-answer testing was superior to restudying, irrespective of retrievability. However, when sub-groups were analyzed regarding retrievability, only items with the lowest retrievability outperformed restudying. The finding that practicing short-answer questions is beneficial in general but most effective when practicing difficult questions seems to be at odds with the finding from the first study where short-answer testing was only beneficial whenever difficulty was lowest. However, as

outlined in Chapter III, the retrieval effort hypothesis can explain these inconsistent findings—at least partially: This finding is in line with the retrieval effort hypothesis. Following this account lower retrievability is associated with higher difficulty and thus successful retrieval of items with lower retrievability can result in more pronounced testing effects. Further explanation for inconsistent findings between studies will be provided in the section Theoretical Implications.

No beneficial effects of multiple-choice questions emerged even when learners had to process all response options in order to answer correctly. Furthermore, this study provided evidence for the absence of direct testing effects as a result of practicing multiple-choice questions constructed in such a manner. These findings are in line with results obtained from Study 1 and with the retrieval effort hypothesis when practicing test questions as outlined in Chapter II.

Summary and Discussion of Study 3

The third study presented in Chapter IV further explored under what conditions multiple-choice testing can have direct beneficial effects in university classrooms. To this end, the moderating role of feedback and practice question difficulty was investigated in a minimal-intervention setting. Furthermore, corrective feedback was employed in a novel way that not only provided the correct answer but enabled learners to revisit course content to obtain feedback in a more effortful and active way.

One main finding from Study 3 was that students infrequently revisited course content irrespective of whether learning content was tested or restudied and that they were unable to obtain the correct information. Apparently, participants were unwilling or unable to obtain corrective feedback. We therefore interpreted performance in the criterial test as the

result of the independent variable practice condition without corrective feedback and the factor retrievability which once again reflected item difficulty.

Another main finding from Study 3 was that there is a testing effect for items with high retrievability. This study is among the first to demonstrate that multiple-choice questions without feedback can foster retention in an applied educational context more than restudying does (for another, see Thomas et al., 2020). The finding that only highly retrievable items benefited from practice tests is in line with the findings from Study 1, the bifurcation model, and the idea that in order to be beneficial, practiced items need to be retrievable from memory or else restudying is the more promising study strategy.

Summary and Discussion of Study 4

The research paradigm employed in Study 4 resembled that of Study 3, however it differed in one central aspect concerning the difficulty of practice test questions: Whereas in Studies 1–3, difficulty of practice test questions was assessed in order to investigate the moderating role on the testing effect, in Study 4, difficulty of multiple-choice questions was adjusted to learners' individual abilities in order to create bespoke multiple-choice questions that should maximize testing effects.

Study 4 compared answering these adaptive multiple-choice questions to non-adaptive multiple-choice questions—that consisted of multiple-choice questions with maximum assistance—as well as to restudy in an authentic educational setting of a university course.

Main findings of this study were that contrary to our assumptions, restudying outperformed both testing conditions and that practicing adaptive questions was equally beneficial as practicing non-adaptive questions. However, in exploratory analyses we

identified conditions that rendered the testing conditions superior to restudying as well as adaptive questions superior to non-adaptive questions.

Theoretical Implications

The first aim of this dissertation was to investigate whether direct testing effects—to date only found under laboratory conditions (e.g., Roediger & Karpicke, 2006a)—are observable and practically relevant in real-world educational settings. It has been shown in all empirical studies that practicing retrieval can have a direct beneficial effect on retention in real-world educational contexts, whenever certain conditions were met.

The second aim of this dissertation was to investigate the role of practice test question difficulty in eliciting direct testing effects. Different accounts have been presented that make predictions about the extent of direct testing effects when question difficulty varies. However, it seems premature to compare different theoretical accounts on the basis of the findings presented in the four studies. For one, the theoretical accounts not only differ in their predictions about the effects of practice test question difficulty but also in their predictions about the effects of success in answering practice test questions. As stated in the introductions, these two variables are naturally confounded: Whenever retrieval difficulty is low, retrieval success is high and vice versa. However, theoretical accounts on the testing effect differ in their assumptions about the effectiveness of different question formats. The bifurcation model does not make any specific predictions in regard to the difficulty of practice test questions, but sees successful retrieval as being always more beneficial than restudying and unsuccessful retrieval, thus expecting the benefits of testing being the most pronounced when retrieval success is highest (Rawson & Zamary, 2019). Other theoretical accounts such as the retrieval effort hypothesis assume that question formats differ in the retrieval effort and thus in the extent of direct testing effects. It was not the aim of this

dissertation to test theoretical accounts on the testing effect and as previously stated, the employed research designs do not allow for such tests, however, findings obtained from studies investigating the effects of different question formats might allow for a tentative conclusion concerning the explanatory value of the theoretical accounts.

The bifurcation model would predict multiple-choice testing being superior because of the higher rates of retrieval success. As illustrated by the first two studies, in direct comparison to restudying contents of a university lecture, short-answer testing produced beneficial effects whereas multiple-choice testing did not, even when multiple-choice questions were constructed in a way that encouraged learners to process all response options. It thus seems likely that these findings that indicate a superiority of short-answer questions over multiple-choice questions can be explained better by the retrieval effort hypothesis.

When recapitulating the findings concerning the moderating role of question difficulty, Study 1 and 3 seem to provide evidence for the bifurcation model because a testing effect was only found when retrievability and thus retrieval success was high. However, given the fact that no feedback was administered subsequent to answering test questions, producing an incorrect answer means never being presented with the correct answer which is also assumed by all other accounts on the testing effect of being inferior to restudying (Kornell & Vaughn, 2016; Rowland, 2014). Following the ideas and terms of the new theory of disuse, unsuccessful retrieval attempts reflect very low retrieval strength which means that the information is not accessible in memory. Not being able to retrieve a piece of information from memory also means that the retrieval route cannot be strengthened and thus no increment in retrieval strength can be obtained by unsuccessful retrieval (R. A. Bjork & Bjork, 1992). The retrieval effort hypothesis is therefore also able to explain results indicating that both higher and lower retrievability of short-answer questions is associated

with more pronounced testing effects. Similar to the bifurcation model this explanation involves the retrieval success rate of all tested items as a crucial variable: Only when an item is correctly recalled from memory it receives a boost in retrieval strength and is thus more likely to be recalled again in a later final test. However, irretrievable items receive no boost in retrieval strength.

Items that have been restudied receive a boost of retrieval strength that is less than the boost obtained by successful retrieval. In the bifurcation model as well as in the retrieval effort hypothesis, the benefit of testing is determined by the amount of successful retrievals because each successful retrieval is associated with an increase in retrieval strength which in turn determines the overall retrievability in the criterial test. To restate a key assumption of both theoretical frameworks, only when the increment in retrieval strength of the tested items surpasses the retrieval strength of restudied items a testing effect emerges.

Additionally, the retrieval effort hypothesis assumes that the boost in retrieval strength rises with the difficulty of the retrieved item (Pyc & Rawson, 2009) thus explaining why difficult questions promote retention more than easier questions do as observed in Study 2.

Additional support for the claim that the retrieval effort hypothesis provides more explanatory value, comes from the finding of the fourth study. Study 4 directly aimed at the optimal trade-off between difficulty and success but also highlighted learners' individual abilities: In order to capitalize on higher difficulties and their potential to increase the benefits of practice tests, learners were required to demonstratet a certain degree of understanding and knowledge of the to-be learned material. Furthermore, varying the difficulty in the condition using adaptive questions and holding difficulty constant in the non-adaptive condition also has theoretical implications. It should be restated that the expected success of retrieval was

similar in both these conditions because even if learners performed badly in the adaptive testing condition, they eventually encountered the lowest difficulty level which was identical to the difficulty level used in the non-adaptive testing condition. Thus, the main difference between these conditions was the possibility to encounter more difficult questions in the condition that presented participants with adaptive questions. The bifurcation model would assume that if any differences between these conditions would occur, it would be the non-adaptive questions outperforming adaptive questions because non-adaptive questions are associated with easier questions resulting in more successful retrieval. In the long run however, a superiority of adaptive testing over non-adaptive testing was found whenever learners were able to answer more difficult questions.

The retrieval effort hypothesis is also able to account for inconsistent findings regarding the beneficial effects of multiple -choice questions in Studies 1–3: Multiple-choice testing was found to elicit direct testing effects in Study 3 whereas no direct testing effects were observable for multiple-choice questions in Study 1 and Study 2. This inconsistency in findings might be explained by a difference in the time intervals that laid between initial learning events and practice tests (i.e., interstudy intervals). In Studies 1 and 2 practice tests occurred at the end of each lecture session capping interstudy intervals to 90 minutes. In Study 3 on the other hand, participants first read assigned chapters in preparation of a course session and then visited course sessions within one week following the course session in which the assigned reading was discussed. This procedure led to interstudy intervals between 2 and 14 days. In increasing the interstudy interval, the information becomes more difficult to be accessed from memory and as a result retrievability is thought to be lower (R. A. Bjork & Bjork, 1992). However it has been shown that this difficulty in retrieving the information from memory can be desirable and that increasing the interstudy interval is associated with

more pronounced testing effects (Pyc & Rawson, 2009). Consequently, it seems likely that answering multiple-choice questions in Study 3 required more effort than in Studies 1 and 2 and this additional effort rendered them beneficial, at least for those questions that led to successful retrieval of the targeted information.

Concerning the second aim of this dissertation, it can be concluded that practice question difficulty is able to explain differing direct testing effects in real-world educational settings and that the retrieval effort hypothesis seems suitable to explain beneficial effects of practice testing in applied university contexts when no feedback is provided.

The third aim of this dissertation was to investigate means to increase the benefits of practicing retrieval in real-world educational settings by artificially changing test question difficulty. Study 4 demonstrated that under some specific conditions, creating a match between learners' abilities and test question difficulty results in more pronounced testing effects compared to standard practice tests.

One important factor moderating both the testing effect and the superiority of adaptive questions was the retention interval. This finding is in line with the desirable difficulty framework, the bifurcation model and findings from laboratory research (R. A. Bjork, 1994; Kornell et al., 2011; Rickard & Pan, 2018; Rowland, 2014; Wheeler et al., 2003). Using the terms of the new theory of disuse (E. L. Bjork & Bjork, 2011; R. A. Bjork & Bjork, 1992), more difficult learning activities are associated with lower retrieval strength because gaining access to the required information is more difficult. However, the lower the retrieval strength of a piece of information is when being correctly recalled from memory, the higher the gain in retrieval strength will be following the recall of that piece of information. Once acquired retrieval strength will gradually decrease over time. In Study 4, restudying required less effort than practicing questions did. Practicing questions furthermore differed in

the required effort because of the operationalization of adaptive and non-adaptive questions: In the adaptive testing condition, participants were initially presented with the most effortful questions and when continuously answering incorrectly, gradually encountered less effortful repetitions of these questions. Participants in the non-adaptive practice condition, on the other hand only encountered the least effortful questions without a chance to increase the needed effort. As a consequence, adaptive questions presumably led to greater gains in retrieval strength and, after longer retention intervals, to sufficient retrieval strength to answer more criterial test questions correctly than the other condition. This explains why longer retention intervals have been found to be linked to diminishing the superiority of restudying over practicing questions and rendering adaptive testing superior to non-adaptive testing.

Another important factor that rendered adaptive testing superior to non-adaptive testing was the participants‘ preparation for the course in terms of completing the reading assignment which probably reflects the best approximation of participants‘ ability levels. The idea of using adaptive practice questions was that to provide participants with solvable but effortful questions by means of altering question difficulty to match the individual learner‘s ability level. We expected adaptive questions to be superior to non-adaptive questions because non-adaptive questions were set to a minimum difficulty whereas adaptive questions allowed for gradual decreases from highest to lowest difficulty, eventually being solved when question difficulty matched the learner‘s ability level. As we found a superiority of adaptive questions only if a learner‘s ability level was high, it seems reasonable to assume that the higher difficulty levels in the adaptive questions—which pose the main difference to non-adaptive questions—were only beneficial whenever a learner’s abilities matched these difficulties and rendered them desirable. In other words, whenever a learner’s abilities were not highest, practicing the more difficult, adaptive questions was equally effective as

practicing the easier, non-adaptive questions. Only when a learne‘s abilities were highest the practice of the more difficult adaptive questions led to challenges both difficult and solvable that are assumed to create the most beneficial testing effects.

Concerning the third aim of this dissertation, it should be concluded that it is possible to capitalize on the assumed connection between question difficulty and testing effect and that under some conditions adaptive testing produced the most beneficial effects on retention.

Practical Implications

The findings presented in this dissertation bear many practical implications for learners and instructors alike. Most importantly, findings presented in this dissertation suggest that practice tests, already popular among learners, are effective in real-world educational settings—however in ways not expected by learners and practitioners: For one, it has been demonstrated that practice tests are a valuable tool to promote retention in real-world educational settings, even when no feedback is provided. For another, findings from two studies presented here suggest that practicing the key contents of a university lecture by means of short-answer questions promotes retention more than answering multiple-choice questions and restudying do. Educators should thus be aware of the beneficial effects of short-answer questions in their everyday teaching and should not only employ practice tests to areas where corrective feedback is possible. Instead, educators could also consider short-answer practice questions as part of asynchronous learning environments, for example as practice questions in books or web-based learning environments. Additionally students should be encouraged to use self-testing, even if no feedback is available or shy the effort to look up a correct answer.

It has furthermore been shown that answering multiple-choice questions was sometimes superior to restudying. However, findings across studies 1–4 were inconsistent to advise the use of multiple-choice questions without feedback. Nevertheless, when feedback can be provided, other works already pointed out that multiple-choice questions can reliably foster retention, presumably via many indirect testing effects (McDaniel & Little, 2019).

Although the methodology employed in the presented studies differs from most studies presented in the literature reviews in terms of feedback and control condition, there is consensus that there is a practically relevant testing effect in real-world educational settings. However, it might not be as ubiquitous as suggested by the many studies which find beneficial effects of testing irrespective of question format, learning content, or educational setting (Moreira et al., 2019).

Apart from demonstrating direct testing effects in existing university settings, it is a merit of studies presented here to provide evidence of these effects and their moderators over time spans that are practically relevant. Studies 3 and 4 investigated the effects of practicing multiple-choice questions after several weeks, Study 1 and 2 investigated retention of practiced learning content at the end of the term. Furthermore, in Study 1, practicing short-answer questions after attending a university lecture was found to be superior not only at the end of the term but even 23 weeks later and thus indicating that this simple intervention had sustainable effects on retention in an actual educational context. Additionally, Study 4 indicated that the effectiveness of practicing questions seems to increase with longer retention intervals. Future research should further investigate this connection between retention intervals and (direct) testing effects.

As initially outlined, students do often prefer restudying to practice testing. Findings presented in this dissertations add to the body of research that this preference is detrimental to

effective learning. Instead, students should be taught the beneficial effects practicing test questions has by itself. It has been shown that instructions to use practice tests ultimately result in more practice testing (Ariel & Karpicke, 2018). Furthermore it has been found that the estimated effectiveness of study strategies is a solid predictor for the use of study strategies (McCabe, 2011) and hence teaching of effective ways to practice testing might result in a higher estimated effectiveness of practice testing and thus more frequent practice testing among learners. Unfortunately, to date little effort has been put into instructing students in the use of effective learning strategies: From a US-sample of textbooks about cognitively based strategies that should maximize student learning and retention, only 44% cover the beneficial effects of testing (Pomerance et al., 2016). Additionally, surveys among instructors indicate that most of them realize the potential of practice tests, however only a fifth of educators think that tests are beneficial independent of corrective feedback and additional study (Morehead et al., 2016). Findings presented in this dissertation suggest that practice testing can be beneficial for retention of once studied learning material even without feedback and additional study when certain conditions are met.

It should be emphasized that although testing is an effective study strategy, it cannot replace learning and teaching. The present research demonstrated instead that thorough understanding and a solidified knowledge are important for the effectiveness of practicing retrieval: Better understanding and knowledge about the learned content will result in higher retrievability rates which has been identified as a crucial pivot point in the occurrence of testing effects in the studies presented here.

Limitations and Directions for Future Research

In the following section, I will discuss the delimitation of studies presented in this dissertation and their combined interpretation.

First of all, many of the conclusions drawn here, were the result of comparisons across multiple studies presented in this dissertation. As each study employed a different methodology and different participants, findings should be carefully interpreted across studies. Although it is a merit of the studies presented here to be among the first to demonstrate direct testing effects in real-world educational contexts, notions concerning the mechanisms and moderators across studies can merely be seen as data driven explanations. Future research should thus investigate mechanisms and moderators of direct testing effects with appropriate designs that allow for adequate conclusions.

In all empirical studies, experimental designs were employed that aimed at maximizing internal and external validity while investigating the research questions. However, some methodological issues are difficult to account for. For one, research documented in this dissertation exclusively investigated learning when teaching psychology courses in university classrooms. Although reviews across all applied studies investigating testing effect find similar results irrespective of the taught domain (Moreira et al., 2019) it remains unclear whether direct testing effects and their moderators are of equal practical relevance in different applied educational settings.

Additional reservations concern the concept of direct testing effects and whether differences between restudying and practice test conditions reflect direct testing effects alone. One might argue that the presented studies not exclusively investigated direct testing effects because in these studies, different conditions might have changed learners' study behavior. As noted by (McDaniel & Little, 2019; Roediger et al., 2011), the experience of being tested might cause a variety of meta-memorial processes that ultimately lead to a better performance in a criterial test. Increased study behavior or more effective studying elicited by the experience of practicing test questions and subsequent feedback often moderates better

performance observable in criterial tests. It thus might seem reasonable to assume that in the presented studies better retention of criterial tests was an indirect effect of testing elicited by testing (all Studies) or feedback (Study 3). However, two methodological aspects of the studies presented in this dissertation are in favor of assuming that indirect testing effects were reduced to a minimum.

The first aspect concerns the design of the presented studies. All presented studies employed a within-subject design that enabled all participants to experience all practice conditions. To explain testing effects as a result of indirect effects such as study behavior would imply, that students participating in this study changed their study behavior after being tested only and that this potential change in study behavior exclusively benefited learning content that has been tested before. Furthermore, research indicates that learners are not always aware of the benefits of different learning strategies. Einstein and colleagues (2012) had students read text passages and measured retention after one week. Following the initially reading of the passages, students either were tested without corrective feedback on the content or reread the same passages. Immediately after each condition, students had to rate how well they thought, the practiced content was learned. Results indicated that students profited more from an additional practice test as compared to an additional reading of the passages. However, the conditions did not differ in respect to the ratings of how well students thought the content was learned. This finding illustrates that when no corrective feedback is given, learners might not even recognize a meta-memorial difference between practice tests and restudying. In line with this finding it might be reasonable to assume that in the presented studies, learners hardly realized any differences on a meta-memorial level because no corrective feedback was obtainable and subsequently learners might not have changed study

strategies. This assumption is in line with the second aspect that contradicts the assumption that findings of the presented studies are mainly due to indirect testing effects.

The second aspect concerns the observation made in Study 3 that participants rarely sought feedback and that requests for feedback were independent of the practice condition. This finding can be interpreted in accordance with the findings by Einstein and colleagues (2012) that testing without corrective feedback does not activate any meta-memorial processes different from restudying. It should additionally be noted that participants in Study 3 were informed about the correctness of their answers unlike participants in the study by Einstein and colleagues. However, even with this additional knowledge, demand for the correct information—in the form of the textbook passage that can answer the practice question—was not sought more often than in the restudy condition where the correct information was already presented as part of the manipulation. I therefore agree with Einstein and colleagues' conclusion that learners are unaware of the potential practice tests have on retention and—based on the findings from Study 3—further extend this conclusion to study decisions in real-world educational contexts. This also affects the interpretability of the findings from all studies presented in this dissertation: Assuming that practice testing is not associated with other meta-memorial processes than restudying is, it seems unlikely that practice tests caused any change in study behavior, study effort or had any motivational effects.

However, it cannot entirely be ruled out that all indirect effects were elicited by practicing tests in the university classroom. This is especially true since there are indirect testing effects that do not require corrective feedback from tests, such as reducing test anxiety by repeated exposure to tests (Roediger et al., 2011). Nevertheless, it seems implausible that these indirect effects of testing that do not require feedback are capable of explaining all

obtained benefits of practice tests without feedback. It thus seems very likely that direct testing effects are practically relevant in applied educational contexts.

An additional reservation about the interpretation of the findings that involve practice test difficulty concerns the operationalization of retrievability and the comparison of more or less retrievable items with restudied items, the item-selection problem. The item-selection problem states that comparison between successful and unsuccessful retrieval attempts is difficult because they are both compared to restudying to determine the effects on retention. However, this comparison is challenging because it is unknown which items would have been retrievable or irretrievable in the restudy condition (Kornell et al., 2009; Pashler et al., 2003). However, as retrievability was operationalized on the item level in the presented studies, the operationalization circumvents the item-selection problem at least in parts. The approach employed in Studies 1–3 calculated item difficulties for individual items from the practice tests and connected these item difficulties to restudy items. By doing so, it was possible to obtain item difficulties for restudied items that can be assumed to reflect the memorability of learned content. This partial circumvention of the item-selection problem is a strong point of the research presented in this dissertation because it strengthens the evidence gathered from investigations of practice test difficulty and how it affects retention.

Conclusion

The aim of this dissertation was to investigate whether practicing test questions alone elicits beneficial effects on retention and thus serves as a learning strategy that can be fruitfully applied to real-world educational contexts. Therefore the applicability and the practical importance of direct testing effects—often found in laboratory research—and their moderators were investigated in existing university courses. A first presentation of relevant research indicated that this subject is of major importance for everyday teaching and learning,

however under-explored in the research literature. The presented findings may serve as initial evidence that practicing multiple-choice as well as short-answer questions has direct beneficial effects on retention of learning material and that the benefits of answering test questions are practically relevant even when no corrective feedback is provided and compared to other learning activities that foster retention.

Furthermore, findings presented in this dissertation underscored the crucial role of difficulty and the connected retrievability play in eliciting testing effects in applied educational contexts without the provision of feedback subsequent to practicing test questions. For one, studies presented here suggest that the test format (i.e., short-answer or multiple-choice questions) which is associated with question difficulty, determines whether practice tests without feedback are beneficial as compared to restudying. For another, findings presented in this dissertation suggest that the benefits of practicing test questions depend on the naturally occurring differences in question difficulty and evidence is provided for the assumption that the beneficial effects of testing can be increased by altering the difficulty of practice questions. Additionally, this dissertation gave some pointers towards theoretical accounts on direct testing effects and their moderators in real-world educational contexts and how the presented findings connect to these accounts.

Evidence accumulated in this dissertation suggests that answering practice questions in real-world educational contexts is beneficial per se and of practical importance, which means that the concept of direct testing effects is not only limited to laboratory research but also applies to existing educational contexts. Furthermore, it has been shown that, compared to restudying, answering practice questions, even without feedback is an effective strategy to promote retention and therefore bears great practical importance because practicing test questions can easily be employed in everyday educational settings. As direct testing effects

have also been found to foster long-term retention of real-world learning contents, practicing test questions can be seen as an effective addition to teaching and learning.

References

Ariel, R., & Karpicke, J. D. (2018). Improving self-regulated learning with a retrieval practice intervention. *Journal of Experimental Psychology: Applied*, *24*(1), 43–56. http://dx.doi.org/10.1037/xap0000133

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher, R. W. Pew, L. M. Hough, & J. R. Pomerantz (Eds.), *Psychology and the real world: Essays illustrating fundamental contributions to society.* (2011-19926-008; pp. 56–64). Worth Publishers.

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). MIT Press.

Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In A. F. Healy, S. M. Kosslyn, & R. M. Shiffrin (Eds.), *From learning processes to cognitive processes: Essys in Honor of William K. Estes* (1992-97939-014; pp. 35–67). Lawrence Erlbaum Associates, Inc.

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

Glover, J. A. (1989). The "testing" phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology*, *81*(3), 392–399.

Kornell, N., Bjork, R. A., & Garcia, M. A. (2011). Why tests appear to prevent forgetting: A distribution-based bifurcation model. *Journal of Memory and Language*, *65*(2), 85–97. https://doi.org/10.1016/j.jml.2011.04.002

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., & Vaughn, K. E. (2016). How retrieval attempts affect learning. In *Psychology of Learning and Motivation* (Vol. 65, pp. 183–215). Elsevier. https://doi.org/10.1016/bs.plm.2016.03.003

Little, J. L., & Bjork, E. L. (2015). Optimizing multiple-choice tests as tools for learning. *Memory & Cognition*, *43*(1), 14–26. https://doi.org/10.3758/s13421-014-0452-8

McCabe, J. (2011). Metacognitive awareness of learning strategies in undergraduates. *Memory & Cognition*, *39*(3), 462–476. https://doi.org/10.3758/s13421-010-0035-2

McDaniel, M. A., & Little, J. L. (2019). Multiple-choice and short-answer quizzing on equal footing in the classroom: Potential indirect effects of testing. In J. Dunlosky & K. A. Rawson (Eds.), *The Cambridge Handbook of Cognition and Education* (1st ed., pp. 480–499). Cambridge University Press. https://doi.org/10.1017/9781108235631.020

Morehead, K., Rhodes, M. G., & DeLozier, S. (2016). Instructor and student knowledge of study strategies. *Memory*, *24*(2), 257–271. https://doi.org/10/gf2zsb

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*, *4:5*. https://doi.org/10/gf2rp4

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

Pashler, H., Zarow, G., & Triplett, B. (2003). Is temporal spacing of tests helpful even when it inflates error rates? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(6), 1051–1057. https://doi.org/10.1037/0278-7393.29.6.1051

Pomerance, L., Greenberg, J., & Walsh, K. (2016). *Learning About Learning* (p. 46). National Council on Teacher Quality. http://www.nctq.org/dmsView/Learning_About_Learning_Report

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

Rawson, K. A., & Zamary, A. (2019). Why is free recall practice more effective than recognition practice for enhancing memory? Evaluating the relational processing hypothesis. *Journal of Memory and Language*, *105*, 141–152. https://doi.org/10/gfxx7n

Rickard, T. C., & Pan, S. C. (2018). A dual memory theory of the testing effect. *Psychonomic Bulletin & Review*, *25*(3), 847–869. https://doi.org/10.3758/s13423-017-1298-4

Roediger, H. L., & Karpicke, J. D. (2006). 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., Putnam, A. L., & Smith, M. A. (2011). Ten benefits of testing and their applications to educational practice. In *Psychology of Learning and Motivation* (Vol. 55, pp. 1–36). Elsevier. 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), 1432–1463. https://doi.org/10.1037/a0037559

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

Wheeler, M., Ewers, M., & Buonanno, J. (2003). Different rates of forgetting following study versus test trials. *Memory*, *11*(6), 571–580. https://doi.org/10.1080/09658210244000414

Acknowledgements

First of all, I would like to thank my advisor, Professor Dr. Tobias Richter for continuously supporting my research ideas and for his constructive feedback at many occasions. The time I was working in his research group was very instructive. Additional thank goes to Professor Dr. Wolfgang Lenhard for co-authoring some of the studies presented here and sharing his views on how students learn and how learning can be optimized in existing university courses.

I also would like to thank my colleagues at the universities of Kassel (most notably from the LOEWE-Project) and Würzburg (especially in the colloquium) for their support, constructive criticism and conversations, that often led to fruitful thoughts and ideas that have been incorporated in this dissertation.

I am also very grateful to the Open Science Foundation and the Sci-Hub, who helped, not only me, to promote the idea of free and open science.

Additional thanks go to Dres. Julia and Simon Schindler for encouragement, guidance, and for sharing PERSONALITY and ATTITUDE.

I owe very special thanks to my parents who encouraged me—willingly or unwillingly—to follow my own ideas and do the things my own way and for their unconditional love, support, and for always listening to my complaints when my way did not work out.

Finally, I am the most thankful to Carla Greving, excellent researcher, my wife, and mother to Theodora and Kornelius, our wonderful children. Listing the countless and manyfold ways she contributed in supporting me in this dissertation would mean adding at least 100 more pages to this dissertation. I however, need to mention her kind and loving support, for always defending the high standards of psychological (open) science, and her

willingness to always discuss theories, results, and methods with me, irrespective of time and place.