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

Sven Greving, Wolfgang Lenhard, and Tobias Richter

University of Würzburg

Author Note

Sven Greving, Department of Psychology IV, University of Würzburg; Wolfgang Lenhard, Department of Psychology IV, University of Würzburg; Tobias Richter, Department of Psychology IV, University of Würzburg.

The research reported in this preregistered report was supported by the Würzburg Professional School of Education. The preregistered report was created in a two-staged process. Stage 1 included introduction, rationale and planned methodology of the study and was peer-reviewed and approved prior to data collection. The approved Stage 1 protocol as well as materials and data are deposited in the repository of the Open Science Framework (https://osf.io/xsd3j/).

Correspondence concerning this article should be addressed to Sven Greving, Department of Psychology IV, University of Würzburg, Röntgenring 10, 97070 Würzburg, Germany. E-mail: sven.greving@uni-wuerzburg.de

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.

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

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 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 1 ABOUT HERE ---

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* distribution 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 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 2 ABOUT HERE ---

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 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 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 3 ABOUT HERE ---

--- FIGURE 1 ABOUT HERE ---

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 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 4 ABOUT HERE ---

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.

Table 1

Possible Combinations of Practice Conditions Across Course Sessions (Sequence of Topics/Conditions Were Counterbalanced Across Participants)

Combination Nr.	Topic		
	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

Table 2

Parameter Estimates for the Models Estimating the Effect of Testing and the Effect of Adaptive Testing realized by two Orthogonally Coded Contrasts (Helmert Coding)

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. 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).

Table 3

Parameter Estimates for the Most Parsimonious Model including Moderators of the Testing Effect.

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
$N_{Participants}$		68		
N_{Items}		60		

Note. Practice condition (contrast-coded: adaptive testing = -1, non-adaptive testing = -1, restudy = 2). Retention interval (centered around M = 17.73).

Table 4

Parameter Estimates for the Most Parsimonious Model including Moderators of the Adaptive Testing Effect.

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
$N_{\text{Participants}}$		68		
N_{Items}		60		

Note. 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).

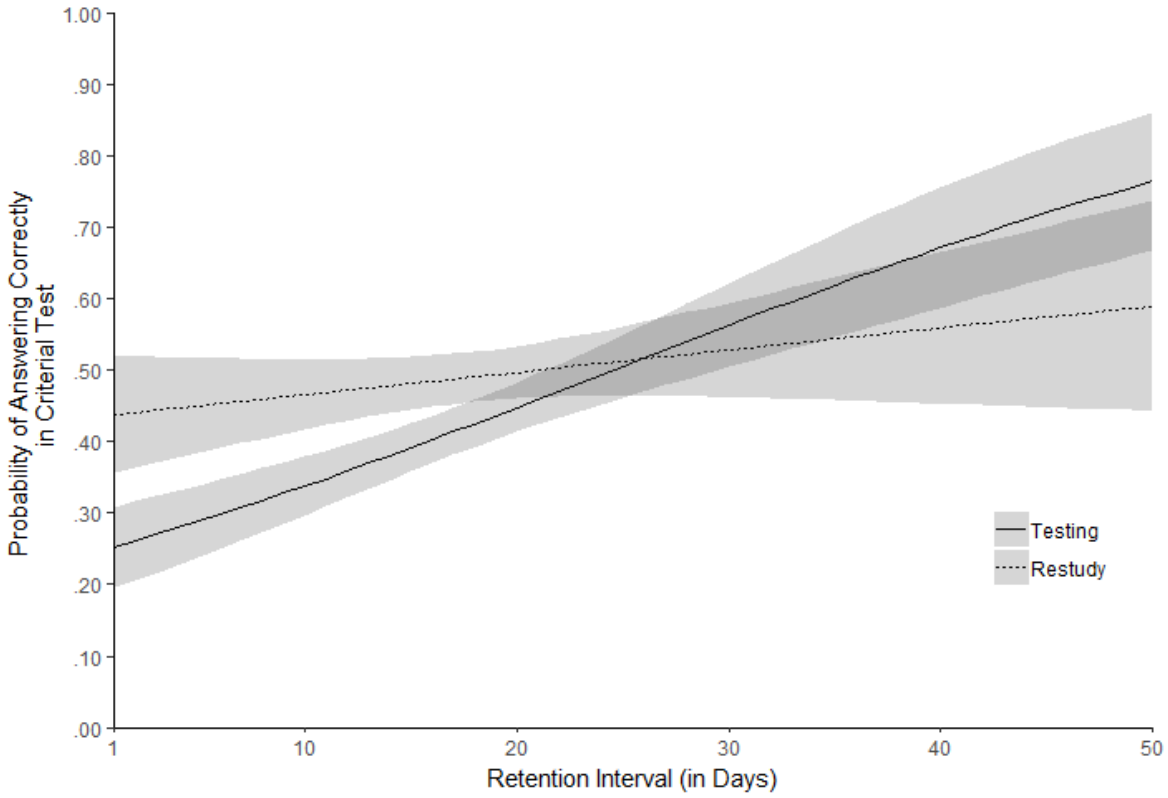

Figure 1. The influence of retention interval on the testing effect. Probability of correct responses in criterial test items (back-transformed from the logits in the GLMM) by retention interval and testing condition (adaptive testing vs. non-adaptive testing). Areas around the graphs indicate standard errors.

Data Availability

The approved Stage 1 protocol as well as materials and data are deposited in the repository of the Open Science Framework (https://osf.io/xsd3j/). Materials used in the study can be made available upon request.