Received: 19 March 2019 | Revised: 18 November 2019 | Accepted: 16 January 2020

DOI: 10.1002/acp.3639

WILEY

RESEARCH ARTICLE

Comparing the effects of generating questions, testing, and restudying on students' long-term recall in university learning

Mirjam Ebersbach | Maike Feierabend | Katharina Barzagar B. Nazari

Department of Psychology, University of Kassel, Kassel, Germany

Correspondence
Mirjam Ebersbach, Department of Psychology, University of Kassel, Hollaendische Str. 36-38, D-34127 Kassel, Germany.
Email: mirjam.ebersbach@uni-kassel.de

Summary

We compared the long-term effects of generating questions by learners with answering questions (i.e., testing) and restudying in the context of a university lecture. In contrast to previous studies, students were not prepared for the learning strategies, learning content was experimentally controlled, and effects on factual and transfer knowledge were examined. Students' overall recall performance after one week profited from generating questions and testing but not from restudying. When analyzing the effects on both knowledge types separately, traditional analyses revealed that only factual knowledge appeared to benefit from testing. However, additional Bayesian analyses suggested that generating questions and testing similarly benefit factual and transfer knowledge compared with restudying. The generation of questions thus seems to be another powerful learning strategy, yielding similar effects as testing on long-term retention of coherent learning content in educational contexts, and these effects emerge for factual and transfer knowledge.

KEYWORDS

desirable difficulties, factual and transfer knowledge, question generation, testing effect, university learning

1 | INTRODUCTION

When students prepare for their exams, they typically restudy the learning material by rereading or rehearsing (Karpicke, Butler, & Roediger, 2009). However, the acquisition of knowledge referring to coherent, complex learning material benefits little from this type of superficial restudying (Callender & McDaniel, 2009), and long-term retention might even be impaired by this strategy (for an overview, see Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013). Long-term retention of curriculum-related material is a central aim in education because prior knowledge facilitates the further acquisition of knowledge and allows knowledge to be applied in a variety of contexts outside formal learning environments, such as when working as a professional. Therefore, identifying learning strategies that promote long-term retention is substantial. We refer to "long-term retention" when retention intervals (i.e., the period between learning and testing) include at least one day (see also Adesope, Trevisan, & Sundararajan, 2017; Agarwal, Karpicke, Kang, Roediger, & McDermott, 2008; Roediger & Karpicke, 2006) in contrast to many laboratory studies in which the learning outcome has often been tested immediately after the learning phase (e.g., Wouters, van Nimwegen, van Oostendorp, & van der Spek, 2013).

1.1 | Desirable difficulties in learning: The testing effect

One branch of learning strategies is predicated on *desirable difficulties*, denoting mechanisms that make the learning process subjectively harder but help learners to retain information in the long run (Bjork, 1994).

One of these desirable difficulties is *testing* by which learners try to answer questions about the learning material during the learning phase before their knowledge is fully consolidated. Testing yields medium to large effects on retention performance in the laboratory and in natural learning contexts (for meta-analyses, see Adesope et al., 2017: Hedges' $g = 0.61$; Rowland, 2014: Hedges' $g = 0.50$).

One explanation for the testing effect is that it promotes retrieval practice when learners try to remember the studied contents during the learning phase (for an overview on retrieval-based learning, see Karpicke, 2017). Retrieval practice has direct effects by strengthening the memory trace through the retrieval attempt and mediated effects by providing feedback to learners about the extent of their learning (Roediger & Karpicke, 2006). In addition, retrieval practice can even enhance the retrieval of other information that is learned after the initial testing phase (for a review, see Pastötter & Bäuml, 2014).

Many studies on the testing effect have focused on the retrieval of facts acquired in the learning phase rather than on transfer effects (for an overview, see Carpenter, 2012). However, Thomas, Weywadt, Anderson, Martinez-Papponi, and McDaniel (2018) reported beneficial and even crossover effects of testing for different knowledge formats in an online learning environment with adult students learning about neuropsychology. Factual questions in the initial testing phase enhanced the final test performance with regard to application knowledge, whereas initial testing with application questions improved the final test performance with regard to factual knowledge. McDaniel, Thomas, Agarwal, McDermott, and Roediger (2013) reported similar transfer effects for the learning of science in middle school but with one exception: Factual questions in the initial testing phase did not improve performance in application questions in the final exam, whereas application questions yielded transfer effects on factual questions in the final exam ($d = 0.34$). Pan and Rickard (2018) conducted a meta-analysis on transfer effects of testing. Transfer was defined relatively broadly, occurring when the cues or required responses (or both) in the initial testing phase and in the final performance tests differed. This definition includes close (e.g., rephrasing information) and far transfer (e.g., drawing new inferences). The meta-analysis revealed a small to medium effect of initial testing on transfer performance ($d = 0.40$). This effect was moderated by several conditions and was negligible when these conditions were not present. Transfer effects were stronger (a) for certain kinds of transfer tasks, for example, for application and inference questions (weaker or even negative transfer effects occurred for questions in which stimulus and response were rearranged compared with the initial test, or for initially presented but untested material), (b) when the initial testing involved the retrieval of broad knowledge, not of isolated concrete facts, and (c) when retrieval was successful in the initial testing phase. Tran, Rohrer, and Pashler (2015) explicitly examined the effect of testing on far transfer by asking participants to making deductive inferences based on premises. Although participants recalled premises to a greater extent when they were initially tested than when they only restudied them, participants' performance in the final test with regard to deductive inferences was not enhanced by initial testing. Given that the majority of studies, focusing on transfer effects of testing, have been conducted in laboratory settings, the authors called for more research on this topic in authentic educational settings. For example, Batsell, Perry, Hanley, and Hostetter (2017) revealed positive effects ($\eta_p^2 = .35$) of quizzing on the performance in the final exam in a university psychology class compared with restudying, and most importantly, this effect also emerged for questions that were not included in the quizzes ($d > 0.59$), which can be conceived as a transfer effect.

Apart from the scarcity of studies investigating the testing effect on transfer knowledge in authentic educational settings, often only the immediate effects of testing on transfer performance were examined in these studies. An exception is the study of Butler (2010) who reported positive effects of initially tested items referring to a short text passage on far transfer in a final test after 1 week ($d = 0.99$). The present study addresses, among other aspects, far transfer by investigating the long-term testing effects on factual and transfer knowledge.

1.2 | Generating questions

Instructing learners to generate questions based on the learning material also yields medium to large effects on comprehension, recall, and problem solving (for an overview, see Song, 2016). Generating questions may stimulate a deeper processing and reflection of the learning material as well as retrieval practice in comparison with restudying. However, in most of the reviewed studies, learners were trained on how to generate questions effectively and practiced this strategy in advance and under supervision. In addition, the learning material involved only short text passages, and only short-term effects were examined.

Bugg and McDaniel (2012), for example, instructed undergraduate students in the laboratory on how to generate either factual or conceptual questions and presented them with examples for each question type. Thereafter, the students were asked to read short text passages on scientific phenomena and to generate questions and answers related to these texts. The generation of questions was compared with rereading. Students had access to the text passages in all conditions (i.e., open-book condition). The final test, including factual and conceptual test questions, took place immediately after the learning phase. The generation of conceptual questions yielded a benefit for conceptual test questions ($\eta_p^2 = .19$) compared with rereading, whereas the generation of factual questions yielded no effect.

Evidence for the effects of the generation of questions on comprehension was accumulated in a meta-analysis across 26 studies in which children and college students were trained in multiple sessions on how to generate questions related to written texts. The analysis yielded medium to large short- and long-term effects on the comprehension of the studied material (Rosenshine, Meister, & Chapman, 1996: $g = 0.61$).

Although many of the studies on question generation were conducted in the laboratory, some studies examined this effect in real learning settings (i.e., in school or at a university). King (1992), for

example, compared the effects of self-questioning, summarizing, and note taking (as the control condition) in the context of videotaped university lectures on sociopolitical themes. Students first received background information and a comprehensive training on self-questioning or summarizing (i.e., a 50-min training phase and four practice phases, 50 min each). Thereafter, students saw the videotaped university lectures and were asked to apply their respective learning strategy. In an immediate comprehension test directly after the last lecture, students in the self-questioning condition and in the summarizing condition outperformed the control group, whereas no difference was found between the first two groups. In the final recall test after one week, self-questioners performed significantly better than summarizers and students in the control group, with no differences between the last two groups (no effect sizes reported).

A similar field study was conducted by King (1994) with fourth and fifth graders within their regular science curriculum. They followed real lessons on the structure and functioning of the body. The children first received an introduction into the respective learning strategy (i.e., generating and answering questions in dyads that either targeted discovering relationships between different concepts within one lesson or relating the lesson content to their prior knowledge) and practiced it during three lessons. Children in a control group were not guided on how to generate questions. Comprehension and knowledge construction, tested one week after the treatment phase, was better in both groups in which children were guided on how to pose questions than in the control group. Thus, with ample training, generating questions in real learning contexts can promote short- and long-term retention in children and adults.

Other studies have been conducted in real learning contexts without training, but they suffer from methodological shortcomings. In the Berry and Chew (2008) study, students were not randomly assigned to the respective learning strategies. Instead, the students decided whether or not they wanted to generate questions about the lecture content. In the Levin and Arnold (2008) study, two experimental question-generation groups were compared but no control group was included in which questions were not generated.

In sum, the extent that generation of questions also yields robust effects on retention performance when learners are not trained remains an open question, as such training and practicing strategies is effortful and time-consuming in real learning contexts. We address this question in the present study by examining the long-term effect of the generation of questions (and answers) in the context of a university lecture without prior training.

1.3 | Studies comparing the effects of testing and the generation of questions

Testing in terms of answering questions generated by others might be conceived as complementary to generating questions oneself. However, questions generated by the learners could also be seen as a form of testing because the previously processed information must be retrieved in the generation phase to formulate adequate questions and answers. An important question is whether both strategies yield similar effects or whether question generation is even superior to testing given that it includes not only responses but also the formulation of the questions.

Studies that compared the effects of testing and question generation were often based on short texts, the method included only short test delays (i.e., a couple of days), and most critically, the conditions were often not comparable with regard to the extent of learning material covered in the testing or question generation condition and the expenditure of time to perform the tasks.

These previous studies yielded contradicting results. A larger benefit of testing compared with generating questions and rereading was reported by Denner and Rickards (1987) for 5th to 11th graders. Weinstein, McDermott, and Roediger (2010) revealed similar benefits from both question generation (d = 0.75) and testing (d = 0.96) compared with rereading in a sample of adult students. Other studies suggested that the generation of questions might be even more effective than answering questions generated by others (e.g., by teachers: Hartman, 1994; Palinscar & Brown, 2009). Foos, Mora, and Tkacz (1994) found a general advantage of students who generated parts of the learning material themselves (including self-generated questions) compared with students who were provided by others with the material (including other-generated questions), g = 0.15. Bae, Therriault, and Redifer (2019) held the learning time constant across conditions, including testing and the generation of questions, and found—in contrast to Foos et al. (1994)—an advantage of testing over the generation of questions in a final test after one week in a sample of students. However, the demands in the learning conditions differed with regard to the tasks included, for example, retrieving all information of the text that could be remembered (i.e., free recall), answering 20 multiple choice questions (i.e., testing), generating an undefined number of exam questions (i.e., question generation), or generating five keywords related to the text (i.e., keywords). Thus, the reported effects could be attributed to these differences between the conditions.

Given the inconsistent findings, the question of whether question generation and testing boost retention to a similar degree compared with restudying when both conditions are comparably manipulated needs to be further investigated. Generating questions could arguably be more favorable than testing because it requires the active reflection of the learning content in search for material that can be reflected in a question, followed by the generation of the corresponding answer. Moreover, the cognitive processing involved in the generation of questions is greater than with testing because with testing, the content is already implied by the question, and only the answer is required to be generated.

These shortcomings when question generation and testing were compared will be addressed in the present study, which fits with the ongoing discussion on teacher- versus student-centered learning (e.g., Kirschner, Sweller, & Clark, 2006). Testing can be conceived as a teacher-centered approach, which addresses content that the teacher believes to be relevant. The generation of questions by students, in contrast, can be conceived as a student-centered approach because

the content reflected in the questions is selected by the learners. Thus, they must discern the relevance and importance of the information when generating questions. Moreover, the strategy also requires the generation of corresponding answers, which might evoke a deeper processing than just answering questions on a test.

1.4 | Open- versus closed-book tests

As outlined in Section 1.1, testing yields robust effects on learning and retention. However, the retrievability of the content in the initial testing phase appears to be a crucial factor for the testing effect (Rowland, 2014). When information is not retrievable in the initial testing phase, it cannot be consolidated by the mere attempt of retrieval. To solve this problem, testing can be conducted with feedback. When learners are given the correct response after having tried to retrieve the response in the initial testing phase, long-term memory is additionally enhanced (Butler & Roediger, 2008). Feedback can be provided either as a formal response to the learners' answers or by offering learners the opportunity to search for the information in their notes or learning material. The latter option is called an open-book test, compared with closed-book tests, where learners are not allowed to use the material in the initial testing phase and do not get explicit feedback, at least until the phase is finished. Open-book tests also reflect more validly what students often do in the frame of their self-regulated learning. Usually, after having memorized new information, students try to recall this information and then look in the learning material when their recall attempt fails.

Agarwal et al. (2008) compared the testing effect in an open-book condition in which learners were allowed to look up the material during initial testing and two closed-book conditions in which learners either completed the initial tests during the learning phase with feedback (i.e., they were provided with the learning material after they completed the initial test and were told to check their answers) or without feedback. Scores in a final test immediately after the learning phase were higher in the open-book condition compared with the two closed-book conditions (d = 1.12). In a second final test after one week, open-book testing outperformed closed-book testing without feedback (d = 0.45), whereas the performance in the open-book condition and the closed-book condition with feedback was similar, and both conditions yielded better retention than simple restudying (ds > 0.87; cf. Nevid, Pyun, & Cheney, 2016 for similar results). Furthermore, the performance in the closed-book condition with feedback was better than in the closed-book condition without feedback in the second final test (d = 0.57). The initial advantage of the open-book test can be attributed to the fact that learners have the chance to correct their memory stores (cf. Gharib, Phillips, & Mathew, 2012). Closed-book tests without feedback, in contrast, do not offer such an opportunity. Thus, learners might potentially recall incorrect information, which is then strengthened by the initial testing. This shortcoming can be prevented by means of feedback in closed-book tests. A recent laboratory study found support for the crucial role of retrievability in the testing effect. Roelle and Berthold (2017) reported an advantage of open-book tests compared with closed-book tests in fostering long-term recall of complex learning material. In contrast, Rummer, Schweppe, and Schwede (2019) reported the opposite finding in a field experiment stretching over multiple seminar lessons, which might be assigned to the fact that students had restudied at home. In sum, the testing effect seems to be more pronounced if retrieval is accompanied by feedback, either by using open-book tests or by closed-book tests with feedback, especially when complex knowledge is tested.

1.5 | The present study

The present study aims at extending empirical findings on the effects of testing and generating questions with regard to the following aspects: We examine the *long-term effects* on the recall of *factual and transfer knowledge* when using this strategy in the context of a *university lecture*. More specifically, we compared a testing condition and a generating questions condition with a restudy condition. The generating questions condition involved *no prior training*. The material indicated the content that students should address when generating questions. This method ensured that this *condition* was *comparable* with the testing condition in which students received questions that addressed the same content. In addition, all students were provided with the same material in the learning phase to enhance the comparability of the conditions. This procedure resulted in an open-book condition for initial testing and the question generation group (see Agarwal et al., 2008). Students were allowed to look up information after they had tried to retrieve the learned content when solving the tasks in the learning phase. Thus, students' performance was less dependent on their retrieval success compared with traditional closed-book conditions. However, the open-book condition when generating questions also increased the comparability of this condition with the restudy condition, which was by nature of its activity an open-book condition, allowing students to correct their long-term memory stores. A final surprise test was administered after one week.[1] The test included both factual and transfer questions to compare whether one type of knowledge benefits more from the different learning conditions. Students were additionally asked how they usually prepare for exams to contrast potential effects of the learning conditions with their learning strategies. Self-testing is not a frequently used learning strategy (Karpicke et al., 2009). Thus, we assumed that generating questions would also not be reported as a frequently used strategy.

We expected that (a) students in the generating questions condition would perform better in the final test after 1 week than students in the testing condition because of the greater generation activity, (b) students in both experimental conditions would outperform students in the restudy condition, and (c) the effects of the generation of questions and testing would emerge for both factual and transfer

knowledge. We additionally explored whether the depth of the questions in the question generation condition was related to students' recall performance.

2 | METHOD

2.1 | Design

The study followed an experimental pre-/post-design. Learning condition (i.e., generating questions, testing, and restudy) served as the between-subjects variable to which students were randomly assigned. All students were tested one week after the learning session to assess their long-term retention. The final test included factual questions that assessed information found in the learning content and transfer questions that assessed students' deeper understanding of the learning content. Final test performance (i.e., proportion correct) was the dependent variable.

2.2 | Participants

Participants were recruited and attended a lecture in developmental psychology. In the experimental learning session at the end of one lecture, 105 students consented to take part (77% female; age: M = 21.8 years, SD = 4.5; 49% psychology students, 43% teacher trainees, and 8% other students). A priori calculations of the sample size required for linear regressions, assuming a medium effect size of learning condition (i.e., f^2 = .15; see Section 1), a power of .90, and including two predictors (i.e., generation of questions and testing, restudy as reference group), yielded an N = 88 (G*Power: Faul, Erdfelder, Buchner, & Lang, 2009). The psychology students were in their first semester and teacher trainees in their third semester. Attending a lecture on developmental psychology before in their studies, addressing the topic covered in the present study, was highly unlikely.

The students were randomly assigned to one of the three learning conditions. They participated voluntarily. However, the learning session and the final test session took place within the course, and the lecture material was relevant for their exam at the end of the semester. As an additional incentive, students who finished both sessions could take part in a lottery.

The final sample, only including students who took part in the learning session and the final test session, consisted of 82 students in total: 30 students in the restudy condition (83% female; age: M = 22.2 years, SD = 5.5), 22 students in the question generation condition (86% female; age: M = 20.2 years, SD = 3.0), and 30 students in the testing condition (70% female; age: M = 21.0 years, SD = 2.5). The decrease in sample size between the lecture that included the experimental learning phase condition and the final test that took place in another lecture can be attributed to the fact that students were not obliged to be present in the lectures. No systematic attrition effect occurred in any of the conditions because the final test was not announced.

2.3 | Material

The lecture was about a topic in the field of developmental psychology (i.e., the development of domain-specific knowledge in infancy and childhood). Usually, students who attend this lecture have not encountered this subject before in their studies. Thus, prior knowledge can therefore be ruled out as an unlikely confound (see also Section 2.1). A paper booklet with demographic questions and an open question about how the student usually prepare for exams (multiple answers were possible) was distributed to all students at the end of the lecture. The booklet also included instructions for the particular learning task and 10 slides of the lecture, which were identical in the three learning conditions (see Supporting information, including the original data, in OSF: https://osf.io/a3w9y/). Relevant words were printed in bold on the slides. In the *generating questions condition*, students were instructed to formulate one exam question in an open response format for the content of each slide and to also provide an answer to the question based on the relevant keywords that were printed in bold. In the *testing condition*, one question per slide was formulated referring to the bold keywords. The students' task was to try to answer the questions first without help and to only look up the answer in the slides if they were not able to provide an answer. In the *restudy condition*, the instruction was to go through all 10 slides and memorize the content. The proportion of questions generated by the learners in the generating questions condition was similar to the proportion of questions answered in the testing condition (i.e., 99% of the requested questions were generated in the generating questions condition and 96% of all questions were answered in the testing condition). However, students in the generating questions condition generated a larger proportion of correct answers in the learning phase (99%) compared with students in the testing condition (83%).

The final surprise test was conducted again within a lecture but this time by means of an internet-based test, accessible via a link by means of smartphones or other electronic devices. The few students who had no electronic device received the tests in a paper–pencil version. The final test was not announced in advance to prevent students from preparing for this test and to rule out self-selection processes. It included 10 factual questions, asking for isolated facts that could easily be derived from the bolded words on the slides (e.g., "Which factors contribute to the development of a Theory of Mind?"). The factual questions were identical to the questions presented to the students in the testing condition in the learning phase. In addition, each final test included 10 transfer questions that assessed students' deeper understanding of the learning content in terms of being able to use it in new contexts (Pan & Rickard, 2018). Transfer questions referred to the same slides as factual questions, but required the application of knowledge beyond the bolded words in the slides such as transferring it to new contexts, making generalizations or inferences (see Appendix A for an example of a slide and the corresponding factual and transfer question; for all questions and answers, see supplementary information in OSF: https://osf.io/wc29h/). Due to a mistake, one transfer question referred to an information that was presented on a slide. This question, which was part

of Item set 2, was therefore excluded from the analyses. To keep the final test sessions short, about half of the students received five factual questions and five transfer questions (i.e., Item set 1), and the other half received the remaining five factual and four transfer questions (i.e., Item set 2). The answers in the final test were scored with 1 to 4 points per question. These scores were summarized across questions and transformed into proportion correct separately for factual and transfer knowledge as well as for the total score. The final test performance of about 50% of the students was rated by a second rater, yielding satisfying interrater reliabilities ranging between .94 and .98.

2.4 | Procedure

Students attended a regular university lecture on the development of domain-specific knowledge as part of their courses. About 20 min before the end of this lecture, students were informed that an extra learning phase would follow that would help them memorize the content of the lecture. In addition, students were informed that there would be different conditions but that all conditions were expected to have positive effects. Thereafter, students were randomly placed in three separate groups to avoid interferences between the conditions, and the materials (i.e., the booklets with the instructions, tasks, and slides) were distributed. After the students finished all the tasks, the booklets were collected. The slides were accessible at all times in all conditions of the learning phase. One week later, at the beginning of the next lecture, the final online test was administered based on information given in the lecture. The final surprise test that took about 20 min was not scheduled in advance to prevent students from preparing for the test. In the final test, students were instructed to respond to the questions without additional help and without communicating. To ensure that cheating did not occur, three to four experimenter assistants supervised the students during the test. All participants were informed about the results of the three learning conditions after the study was finished but before the exam took place. This procedure was implemented to counteract any disadvantages due to the imposed strategy, especially for students in the restudy condition, which is known to be less effective compared with testing (e.g., Roediger & Karpicke, 2006). Thus, all students had the chance to use the most effective strategy to boost their performance for the exam.

3 | RESULTS

3.1 | Preliminary analyses

First, we analyzed the learning strategies that the students had reported before the learning session, which they typically use when preparing for exams.[2] More than one strategy could be mentioned. Individual learning strategies were categorized into (a) restudying the material, (b) active summarizing (e.g., writing notes, summaries, and note cards), (c) elaboration (e.g., visualization, working examples, and consulting further resources), (d) self-testing and explaining to others, (e) generating questions, (f) miscellaneous (interrater reliability: Cohen's kappa = .91). Active summarizing strategies were reported most frequently (130),[3] followed by restudying strategies (116). Clearly, strategies that included testing (50), elaboration (19), and other strategies (15) were reported less frequently. Generating questions was reported only once.

We then checked in advance whether the final test performance varied as a function of the study course (i.e., psychology, education, and other courses). No differences were found between the students from different study courses, $p = .27$. Therefore, the data were collapsed across these groups.

Finally, we checked whether the assignment of students to the parallel item sets was balanced in each condition. A simple cross tabulation revealed that somewhat more than half of the students (63%) received Item set 2 in the restudy condition, and somewhat more than half of the students received Item set 1 in the generating questions condition (59%) and the testing condition (57%). To control for this not perfect distribution of item sets across the conditions, the *item set* variable was included in all of the following models.

3.2 | Testing the hypotheses

A linear regression model was computed in R (R Core Team, 2017; RStudio Team, 2016) to test our first two hypotheses that (a) students in both experimental conditions would outperform students in the restudy condition and (b) students in the generating questions condition would perform better in the final test after one week than students in the testing condition. Packages used for data preparation and analyses were dplyr (Wickham, Francois, Henry, & Müller, 2017) and emmeans (Lenth, 2018)[2].

In the linear regression model, *learning condition* (three levels: restudy, generating questions, and testing; restudy as reference group) and the control variable *item set* (1 or 2) were included as predictors. The criterion variable was the *overall final test performance*, measured as proportion correct across factual and transfer knowledge items. The results are shown in Table 1, including the unstandardized regression coefficients that can be interpreted as percentage points by which one group differed from the reference group (restudy). Given that the dependent variable (i.e., proportion correct) could range between 0 and 1, a value of .20, for example, would indicate that the students in that condition scored 20 percentage points higher in the final test than the reference group. The analyses revealed significant positive effects for the generating questions and testing compared with the restudy condition (see Table 1: Model 1 and Figure 1 for descriptive statistics). Students in both experimental conditions (i.e., generating questions and testing) scored on average 11 percentage points higher on the final test compared with students in the restudy condition. No significant difference was found between generating questions and testing, $p = .93$.

TABLE 1 Linear regression models predicting test performance one week after learning

	Dependent variable Overall test performance (Model 1)	Factual knowledge (Model 2)	Transfer knowledge (Model 3)
Intercept	0.45*** (0.04)	0.58*** (0.05)	0.37*** (0.05)
Question generation (ref.: restudying)	0.11* (0.05)	0.11 (0.06)	0.11 (0.06)
Testing (ref.: restudying)	0.11* (0.05)	0.13* (0.06)	0.07 (0.05)
Item set 2(ref.: item set 1)	−0.15*** (0.04)	−0.30*** (0.05)	0.12* (0.05)
R^2 (adj.)	.21***	.38***	.06

Note: Models include unstandardized regression coefficients; standard errors in parentheses. Ref. indicates the reference category against which the target category was tested; *item set* serves as control variable. N = 82.
***p < .001.; *p < .05.

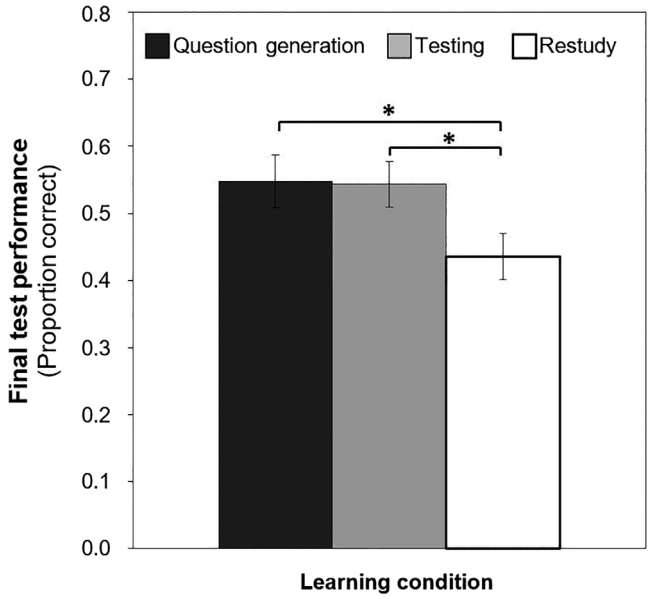

FIGURE 1 Final test performance (proportion correct) one week after the learning session, separately for each learning condition

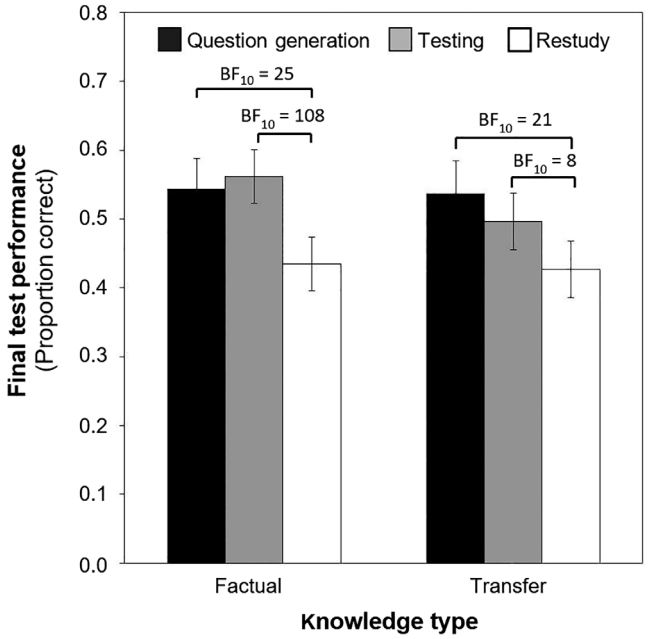

FIGURE 2 Final test performance (proportion correct) 1 week after the learning session, separately for factual and transfer knowledge and for each learning condition. BF_{10} indicates how much more likely, based on the presented data, that the respective experimental learning condition compared with the restudy condition has a positive effect than a negative effect

To test our third hypothesis that the effects of question generation and testing would emerge for factual and transfer knowledge, two additional models were computed, one for each knowledge type (i.e., factual and transfer) with the same independent variables as in the first model (see Table 1 and Figure 2 for descriptive statistics). In the second model with *factual knowledge* as dependent variable, a significant positive effect of testing was found compared with restudying (13 percentage points difference), but no significant effect of generating questions was found compared with restudying, p = .08. The difference between generating questions and testing was also not significant, p = .76. In the third model with *transfer knowledge* as dependent variable, the effects were not significant for generating questions (p = .09) and testing (p = .24) compared with restudying. Furthermore, generating questions and testing did not differ, p = .53.

In order to alternatively check the insignificant results, we reanalyzed the respective regression models by means of Bayesian analyses, using the R package brms (Bürkner, 2017). The Bayesian approach is more advantageous for small sample sizes and allows to test null effects. In contrast to classical inferential statistics, it provides relative evidence for the null or alternative hypothesis in the form of a Bayes factor instead of a binary decision. We report the 95% credible interval for each reported effect, which indicates the range of values that are most likely for the respective effect. Additionally, based on the distribution of the possible parameter values, the Bayes factor BF_{10} can be used to express the likelihood ratio that the alternative hypothesis is correct and the likelihood that the null hypothesis is correct, given the data. For example, a Bayes factor of BF_{10} = 10 would indicate that the alternative hypothesis is 10 times more likely than the null hypothesis (the complement Bayes factor BF_{01} is used to

express the likelihood ratio that the *null* hypothesis is correct and the likelihood that the *alternative* hypothesis is correct). Improper flat priors over the reals were used for the analyses, which means that the prior distributions had little influence on the results and were instead mainly driven by the data (Bürkner, 2017; Kruschke, 2013).

The Bayesian regression analysis with the same variable structure as the models described above for the *overall final test performance* confirmed that no evidence exists for a difference between generating questions and testing (95% credible interval for the effect of testing compared with generating questions from −0.11 to 0.10, BF_{10} = 1). In contrast to the nonsignificant effect of generating questions on *factual knowledge* compared with restudying, the Bayesian model provided strong evidence for a positive effect of generating questions compared with restudying (95% credible interval from −0.01 to 0.23, BF_{10} = 25). That is, although the effect was not significant in the more traditional frequentist approach, the Bayesian analysis suggests that generating questions compared with restudying is more likely to have had a positive effect on performance than a negative or no effect. In addition, the nonsignificant difference between testing and generating questions on factual knowledge was confirmed by the Bayesian model (95% credible interval for the effect of testing compared with generating questions from −0.10 to 0.14, BF_{10} = 0.6). Finally, the more traditional analysis revealed that both effects of generating questions and testing on *transfer knowledge* compared with restudying were not significant. However, the Bayesian model indicated strong evidence for a positive effect of generating questions of about 11 percentage points compared with restudying (95% credible interval from −0.02 to 0.24, BF_{10} = 21) and moderate evidence for a positive effect of testing of about 7 percentage points compared with restudying (95% credible interval from −0.05 to 0.19, BF_{10} = 8). The model also indicated no difference between the effects of generating questions and testing (95% credible interval for the effect of testing compared with generating questions from −0.17 to 0.08, BF_{10} = 3) (see Figure 2).

In sum, the more traditional analyses showed that generating questions and testing—compared with restudying—improve the overall final test performance in the long run and that a similar effect also emerges at least for testing when factual knowledge is analyzed separately, whereas no such effects emerged for transfer knowledge. However, Bayesian analyses indicated positive effects of generating questions and testing compared with restudying on both factual and transfer knowledge, although the effects tend to be smaller than the effects on the overall final test performance (which might explain why they did not reach statistical significance in the more traditional approach). Furthermore, question generation did not outperform testing.

We additionally analyzed whether the depth of the generated questions affected the effect of generating questions. The depth of the questions was evaluated by two raters[4] according to the scoring scheme adapted from Berry and Chew (2008). When factual knowledge or the definition of a concept was addressed in a question, question depth was scored as 1; when the question addressed application-related transfer knowledge, it was scored as 2; and when the question required a deeper conceptual analysis or the integration with other knowledge domains, it was scored as 3. Interrater reliability was $r(290) = .89$. Mean question depth did not vary much between participants (min: 1, max: 1.5, $M = 1.22$, $SD = 0.142$). Moreover, no significant correlation was found between question depth and the final test performance in the generating questions condition, neither when the overall final test performance was considered nor when factual and transfer knowledge were considered separately, $ps > .49$.

4 | DISCUSSION

4.1 | Summary of the findings

The aim of this field study was to examine and compare the effects of generating questions, testing, and restudying on final test performance that addressed the content of a university lecture. In contrast to previous studies, we specifically investigated long-term effects on factual and transfer knowledge, and students were not trained in advance on how to generate questions effectively. In addition, we made the conditions maximally comparable with rule out confounding effects (e.g., by differences concerning the contents covered in the different learning conditions).

Students who generated questions and answers performed similarly well as students who answered experimenter-generated questions (i.e., testing) on their overall performance in the final test after one week. An important finding is that students in both conditions performed significantly better than students who had only restudied the material. The positive effects of question generation and testing compared with restudying were also confirmed when factual and transfer knowledge were analyzed separately. Although a traditional frequentist approach revealed no significant effects of question generation on factual knowledge and of testing and question generation for transfer knowledge, additional Bayesian analyses suggested strong evidence that question generation yielded positive effects on factual and transfer knowledge compared with restudying and moderate evidence that testing yielded a positive effect on transfer knowledge. The depth of the generated questions was not related to students' final test performance.

Our results show that generating questions in an open-book format is—like testing—a powerful learning strategy in real learning contexts that may help students enhance and consolidate their knowledge over longer periods of time compared with restudying. This finding is important because one central aim of education is to promote the long-term retention of knowledge so that it can be applied in different contexts. In addition, long-term retention supports the acquisition of new knowledge by facilitating its assimilation with prior knowledge.

How can the effect of question generation be explained? In general, it is assumed that generating questions stimulates a deeper elaboration of the learning material and a *deeper processing* (King, 1992; Song, 2016). Furthermore, *rephrasing* might be a plausible mechanism explaining the effect (Doctorow, Wittrock, & Marks, 1978; Wittrock,

1974). The extant literature provides ample evidence showing that rephrasing or paraphrasing is an effective tool for enhancing the processing, comprehension, and recall of the paraphrased information (e.g., Bui, Myerson, & Hale, 2013; Hagaman, Casey, & Reid, 2012; Rosenshine & Meister, 1994; Wammes, Meade, & Fernandes, 2017). Rephrasing establishes representational variability of the learning content and therewith generates multiple memory traces to retrieve this content. This assumption is related to the encoding variability hypothesis, stating that retrieval of information is facilitated when it is encoded in multiple ways or by different encoding strategies (Estes, 1950; Glenberg, 1979). Moreover, rephrasing can also be conceived as a *generative activity* because new wording is created. Previous research on the generation effect as another desirable difficulty in learning (Bjork, 1994) has shown that information not only enhances memory when larger parts or whole words from the learning material are generated by the learner (for a meta-analysis see Bertsch, Pesta, Wiscott, & McDaniel, 2007) but also when only single letters of the words are generated or switched by the learner (Donaldson & Bass, 1980; Nairne & Widner, 1987). Thus, even a slight generation activity can be effective. Arguably, the generation of new words is not necessary to stimulate a generation effect. The effect can be invoked, for example, by setting words in a different order in which only details have to be changed. However, further research is needed to clarify how rephrasing and generation contribute to the positive effect of generating questions. Our results also suggest that retrieval practice might not be the essential factor constituting the effect of question generation because retrieval practice was the weakest learning strategy in the present study given the open-book format (cf. Agarwal et al., 2008).

We also demonstrated that generating questions has a significant impact on knowledge acquisition even when learners were not prompted or trained in advance on how to generate questions effectively as in previous studies (e.g., King, 1992) and when questions and answers were not evaluated by the instructor or others afterwards (Song, 2016). Thus, the application of this strategy in educational practice requires little effort and boosts learning. For example, teachers could instruct students to generate exam questions during the lecture from the perspective of the teacher. To provide an incentive for students, the lecturer could tell the students that selected questions would be included in the exam. This practice has been informally reported by several lecturers who taught university courses.

Generating questions yielded similar effects as testing in our study, and both conditions outperformed simple restudying. These effects also became evident for the two different knowledge types (i.e., factual and transfer knowledge) when Bayesian analyses were applied. Given that the open-book format clearly limited retrieval practice in the question generation and testing conditions, other mechanisms could have contributed to the positive effects, as discussed earlier. One advantage of open-book formats is that learners can consolidate correct knowledge by looking up the material (e.g., Agarwal et al., 2008) in contrast to a closed-book condition in which they might not recall the information if it is too complex, or they might recall the wrong information (cf. Roelle & Berthold, 2017; van Gog & Sweller, 2015). We showed that an open-book condition is not only effective in combination with testing but also in combination with the generation of questions. Moreover, an open-book condition corresponds to the typical approach of learners when they test themselves in a self-regulated learning environment (e.g., Kornell & Son, 2009; Wissman, Rawson, & Pyc, 2012).

The finding that testing did not outperform the generation of questions contradicts the findings of Bae et al. (2019) and Foos et al. (1994). However, in these studies, the generation of questions condition was not fully comparable with the testing condition with regard to the number of questions and the content. As a result, students in previous studies could have generated less questions, or multiple questions based on the same aspect of the learning material, or questions that only addressed easily comprehensible aspects, thereby failing to exhibit similar effects as in the testing condition that covered a broader range of learning content. We overcame this problem by prescribing the number of questions to be generated and the content that should be addressed in the generating questions to be able to compare it with the testing condition. The instructions, for example, to form questions based on the bolded words, were followed by the students. One possible critique of our method could be that students in the testing condition had the advantage by receiving the same test items on the final test as they had in the learning phase. Thus, they could have been more familiar with the final test questions than students in the generating questions condition, which in turn could have leveled out a potential advantage of question generation. We tried to rule out this effect by prescribing the terms that should be included when generating questions (i.e., bolded words on the slides). These terms were also included in the questions of the testing condition. Thus, both conditions were comparable in terms of the core content of the questions in the learning phase.

Nonetheless, our finding that testing and question generation without prior training yielded similar effects in a real learning context is promising. Our results suggest that the two learning strategies, which are both clearly more effective than simple restudying, can be recommended by (university) teachers to learners and can also be recommended for (self-regulated) learning. Furthermore, the effects of question generation and testing emerged for factual and transfer knowledge, confirmed by the Bayesian analyses. The finding for testing is in line with results of the meta-analysis of Pan and Rickard (2018) who reported transfer effects of testing, which were particularly strong for application and interference questions that fall in the same category as our transfer questions. However, we also showed that question generation may positively affect factual and transfer knowledge, despite the small effects when the two knowledge types were analyzed separately.

Apart from its positive effects, generating questions—like testing—is also an effortful strategy for learners. The learning process activated by generating questions and testing is more difficult than restudying, and their lack of use can be inferred from our findings. We also observed unsystematically during the experiment that students in the generating questions condition and in the testing condition took slightly longer than students in the restudy condition (cf. Weinstein et al., 2010). The extra time makes sense because a more intensive

examination and a deeper elaboration of the learning content, induced by the generation of questions and testing, takes more time than when only trying to memorize content (Endres, Carpenter, Martin, & Renkl, 2017). However, our study did not explicitly control for time spent on the material despite the fact that the learning period was fixed in all conditions. In a self-paced environment, one could test whether time on task could have contributed to the beneficial effects of testing and generating questions (see also Hoogerheide, Staal, Schaap, & van Gog, 2019). This issue should be addressed in future research to separate qualitative effects of the learning condition from simple quantitative effects of study time. In addition, it might be instructive to assess cognitive load to further explain the revealed effects.

Finally, the depth of the generating questions was not related to students' final test performance in this condition, which is in line with the findings of Berry and Chew (2008). Nevertheless, other studies found that students performed better in recall tests when they were trained to generate cognitively challenging questions (e.g., Bugg & McDaniel, 2012; Levin & Arnold, 2008). The null finding of question depth in our study might be due to the fact that (a) the generating questions mostly addressed superficial facts rather than requiring conceptual analyses or inferences (i.e., low question depth in general) and that (b) question depth varied little in the student sample. The finding of a strong effect of generating questions is astonishing, given that students were not instructed in advance on how to generate questions and given that the questions were not very elaborate. Thus, the generation of questions might be a rather effective strategy independent of prior training.

4.2 | Limitations, future directions, and implications

Several aspects of our study might limit the generalizability of the results and suggest further research. The first limitation is that testing and generating questions were based on an open-book format, that is, students were allowed to search for information in the material to generate questions and answers. Comparing an open- and a closed-book condition on the effects of generating questions and testing would be informative. Studies have shown that the testing effect is stronger in an open-book condition, especially for complex learning material (Agarwal et al., 2008; Roelle & Berthold, 2017). Open-book testing is nevertheless not a common technique in formal instruction, even if it reduces test anxiety when applied in exams (e.g., Gharib et al., 2012). However, the self-paced study time of students might decrease because the students are expecting an open-book exam (Agarwal & Roediger, 2011).

Interestingly, students in the testing condition of the present study generated less correct answers during the learning phase (83%) than the students in the question generation condition (99%). Thus, the activity of generating appropriate questions and answers in an open-book condition might stimulate students to a stronger degree than the activity of testing to look up the material. This difference in practice performance might have also accounted at least to some degree for the fact that testing did not outperform question generation in the present study. To tease out potentially smaller differences between testing and question generation, we recommend replicating our method with a larger sample size.

A second limitation refers to the design of the material in which relevant aspects were printed in bold. In self-regulated learning, students often must identify the relevant aspects in the learning material before memorizing them. Thus, examining the effects of testing and question generation in future studies with material that does not indicate the relevant aspects in advance would be informative.

Third, the topic used in our study was rather specific (i.e., the development of domain-specific knowledge). Future studies could investigate whether the results can be replicated with other topics, for example, with statistics, which is more abstract and requires procedural and descriptive skills. In addition, the retention phase can be extended. The testing effect is known to become stronger over longer intervals between learning and testing (Roediger & Karpicke, 2006). Investigating how the effect of question generation evolves across longer intervals would also be informative.

A fourth limitation is common to many studies that sample university students. The psychology students and teacher trainees in the present study had undergone a rigid selection process to get a place at the university, which was primarily based on their final high school certificate grade (i.e., the German *Abitur* as the main criterion used for the *numerus clausus* policy). In contrast to university students, poor learners could conceivably have difficulties when generating questions. Thus, the generalization of the results using other groups of students or learners (e.g., pupils or high school students) in real learning contexts should be explored in further studies. In addition, the fluctuation and dropout rate was fairly high in our study. This fluctuation is normal for the university where this study took place as students are not obliged to attend the courses. Hence, our method mirrored characteristics of a real learning situation. A constant sample across measurements, however, could provide more valid results. Nevertheless, we believe that our study makes an important contribution to the literature on the efficacy of generating questions and testing in applied settings, such as in university lectures, and some of the shortcomings in comparison with lab studies can also be conceived as strengths in terms of the applicability in real-life learning contexts.

Related to the discussion about the sample is the question of prior knowledge. Despite the unlikelihood that students had previously been introduced into the development of domain-specific knowledge (see also Sections 2.1 and 2.3), prior knowledge on the topic cannot be fully ruled out. However, if some students had prior knowledge of the topic, these students would have been distributed equally across the learning conditions because of random sampling and thus should not have biased the main effect of the condition. In future studies, working with a larger sample might be fruitful, even if an a priori sample size calculation yields a similar sample size as we had. The effects of the learning conditions in the present study were fairly small when the knowledge types were considered separately. A larger sample size might yield stronger effects with more traditional statistical analyses.

Finally, the moderate performance of students in the final surprise test warrants a closer look (see Figure 1). This performance might be due to the fact that this test was unannounced, and the final exam took place about one month after the end of the study. Thus, most students might not have started to study the lecture content. In addition, the performance level shows that the learning content was rather complex and was not easy to learn from a single lecture, even when additional learning opportunities are given.

In sum, although further research is necessary to replicate and extend our findings, generating questions is a learning strategy as powerful as testing in real learning contexts, which requires no extensive training or prompts, and might strengthen students' long-term recall of factual and transfer knowledge. Given its task characteristics and requirements and the fact that students hardly use this strategy spontaneously when learning, the generation of questions might be conceived as a further learning strategy related to desirable difficulties in learning (Bjork, 1994).

ACKNOWLEDGMENTS

We thank Karina Senftner for her support in the data collection and scoring and Mike Cofrin for proofreading the manuscript.

CONFLICT OF INTEREST

Hereby we declare that there is no conflict of interest.

ENDNOTES

[1] The identical final test was repeated four weeks after the learning phase. However, as testing time was manipulated within-subjects, a general testing effect in all conditions cannot be ruled out. Therefore, the results of this second test are not reported here but can be inspected as Supporting information: https://osf.io/a3w9y/.

[2] The data that support the findings of this study are available from the corresponding author upon reasonable request.

[3] These frequencies could exceed total sample size when individuals mentioned two or more strategies that were assigned to the same superordinate strategy.

[4] One rater was one of the authors, and the second rater was a student assistant who coded the questions according to a predefined scoring scheme.

DATA AVAILABILITY STATEMENT

Supplementary material, including the original data, in OSF: https://osf.io/a3w9y/

ORCID

Mirjam Ebersbach https://orcid.org/0000-0003-3853-4924

Katharina Barzagar B. Nazari https://orcid.org/0000-0003-4909-272X

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

Agarwal, P. K., Karpicke, J. D., Kang, S. H. K., Roediger, H. L., & McDermott, K. B. (2008). Examining the testing effect with open- and closed-book tests. *Applied Cognitive Psychology*, *22*, 861–876. https://doi.org/10.1002/acp.1391

Agarwal, P. K., & Roediger, H. L. (2011). Expectancy of an open-book test decreases performance on a delayed closed-book test. *Memory*, *19*, 836–852. https://doi.org/10.1080/09658211.2011.613840

Bae, C. L., Therriault, D. J., & Redifer, J. L. (2019). Investigating the testing effect: Retrieval as a characteristic of effective study strategies. *Learning and Instruction*, *60*, 206–214. https://doi.org/10.1016/j.learninstruc.2017.12.008

Batsell, W. R., Perry, J. L., Hanley, E., & Hostetter, A. B. (2017). Ecological validity of the testing effect: The use of daily quizzes in introductory psychology. *Teaching of Psychology*, *44*, 18–23. https://doi.org/10.1177/0098628316677492

Berry, J. W., & Chew, S. L. (2008). Improving learning through interventions of student-generated questions and concept maps. *Teaching of Psychology*, *35*, 305–312. https://doi.org/10.1080/00986280802373841

Bertsch, S., Pesta, B. J., Wiscott, R., & McDaniel, M. A. (2007). The generation effect: A meta-analytic review. *Memory & Cognition*, *35*, 201–210. https://doi.org/10.3758/BF03193441

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

Bugg, J. M., & McDaniel, M. A. (2012). Selective benefits of question self-generation and answering for remembering expository text. *Journal of Educational Psychology*, *104*, 922–931. https://doi.org/10.1037/a0028661

Bui, D. C., Myerson, J., & Hale, S. (2013). Note-taking with computers: Exploring alternative strategies for improved recall. *Journal of Educational Psychology*, *105*, 299–309.

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*, 1–28. https://doi.org/10.18637/jss.v080.i01

Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *36*, 1118–1133. https://doi.org/10.1037/a0019902

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

Callender, A. A., & McDaniel, M. A. (2009). The limited benefits of rereading educational texts. *Contemporary Educational Psychology*, *34*, 30–41. https://doi.org/10.1016/j.cedpsych.2008.07.001

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

Denner, P. R., & Rickards, J. P. (1987). A developmental comparison of the effects of provided and generated questions on text recall. *Contemporary Educational Psychology*, *12*, 135–146. https://doi.org/10.1016/S0361-476X(87)80047-4

Doctorow, M., Wittrock, M. C., & Marks, C. (1978). Generative processes in reading comprehension. *Journal of Educational Psychology*, *70*, 109–118. https://doi.org/10.1037/0022-0663.70.2.109

Donaldson, W., & Bass, M. (1980). Relational information and memory for problem solutions. *Journal of Verbal Learning & Verbal Behavior*, *19*, 26–35.

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

Endres, T., Carpenter, S., Martin, A., & Renkl, A. (2017). Enhancing learning by retrieval: Enriching free recall with elaborative prompting. *Learning and Instruction*, *49*, 13–20. https://doi.org/10.1016/j.learninstruc.2016.11.010

Estes, W. K. (1950). Toward a statistical theory of learning. *Psychological Review*, *57*, 94–107. https://doi.org/10.1037/h0058559

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*, 1149–1160.

Foos, P. W., Mora, J. J., & Tkacz, S. (1994). Student study techniques and the generation effect. *Journal of Educational Psychology*, *86*, 567–576. https://doi.org/10.1037//0022-0663.86.4.567

Gharib, A., Phillips, W., & Mathew, N. (2012). Cheat sheet or open-book? A comparison of the effects of exam types on performance, retention, and anxiety. *Psychology Research*, *2*, 469–478. https://doi.org/10.17265/2159-5542/2012.08.004

Glenberg, A. M. (1979). Component-levels theory of the effects of spacing of repetitions on recall and recognition. *Memory & Cognition*, *7*, 95–112.

Hagaman, J. L., Casey, K. J., & Reid, R. (2012). The effects of the paraphrasing strategy on the reading comprehension of young students. *Remedial and Special Education*, *33*, 110–123. https://doi.org/10.1177/0741932510364548

Hartman, H. J. (1994). From reciprocal teaching to reciprocal education. *Journal of Developmental Education*, *18*, 2–8.

Hoogerheide, V., Staal, J., Schaap, L., & van Gog, T. (2019). Effects of study intention and generating multiple choice questions on expository text retention. *Learning and Instruction*, *60*, 191–198. https://doi.org/10.1016/j.learninstruc.2017.12.006

Karpicke, J. D. (2017). Retrieval-based learning. A decade of progress. In J. H. Byrne (Hg.): *Learning and memory. A comprehensive reference* (2nd edition. 487–514). Oxford, UK: Academic Press.

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*, 471–479. https://doi.org/10.1080/09658210802647009

King, A. (1992). Comparison of self-questioning, summarizing, and notetaking-review as strategies for learning from lectures. *American Educational Research Journal*, *29*, 303–323. https://doi.org/10.3102/00028312029002303

King, A. (1994). Guiding knowledge construction in the classroom: Effects of teaching children how to question and how to explain. *American Educational Research Journal*, *31*, 338–368. https://doi.org/10.3102/00028312031002338

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist*, *41*, 75–86.

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

Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, *142*, 73–603. https://doi.org/10.1037/a0029146

Lenth, R. V. (2018). emmeans: Estimated Marginal Means, aka Least-Squares Means. R package version 1.3.0. https://CRAN.R-project.org/package=emmeans

Levin, A., & Arnold, K.-H. (2008). Fragen stellen, um Antworten zu erhalten – oder Fragen generieren, um zu lernen? *Zeitschrift für Pädagogische Psychologie*, *22*, 135–142. https://doi.org/10.1024/1010-0652.22.2.135

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

Nairne, J. S., & Widner, R. L., Jr. (1987). Generation effects with nonwords: The role of test appropriateness. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *13*, 164–171.

Nevid, J. S., Pyun, Y. S., & Cheney, B. (2016). Retention of text material under cued and uncued recall and open and closed book conditions. *International Journal for the Scholarship of Teaching and Learning*, *10*, Article 10. https://doi.org/10.20429/ijsotl.2016.100210

Palinscar, A. S., & Brown, A. L. (2009). Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. *Cognition and Instruction*, *1*, 117–175. https://doi.org/10.1207/s1532690xci0102_1

Pan, S. C., & Rickard, T. C. (2018). Transfer of test-enhanced learning: Meta-analytic review and synthesis. *Psychological Bulletin*, *144*, 710–756. https://doi.org/10.1037/bul0000151

Pastötter, B., & Bäuml, K.-H. (2014). Retrieval practice enhances new learning. The forward effect of testing. *Frontiers in Psychology*, *5*, 286. https://doi.org/10.3389/fpsyg.2014.00286

R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org

Roediger III, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*, 249–255. doi: 0.1111/j.1467–9280.2006.01693.x

Roelle, J., & Berthold, K. (2017). Effects of incorporating retrieval into learning tasks: The complexity of the tasks matters. *Learning and Instruction*, *49*, 142–156. https://doi.org/10.1016/j.learninstruc.2017.01.008

Rosenshine, B., & Meister, C. (1994). Cognitive strategy instruction in reading. In D. A. Hayes & S. A. Stahl (Eds.), *Instructional models in reading* (pp. 85–107). Hillsdale, NJ: Erlbaum.

Rosenshine, B., Meister, C., & Chapman, S. (1996). Teaching students to generate questions: A review of the intervention studies. *Review of Educational Research*, *66*, 181–221. https://doi.org/10.3102/00346543066002181

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

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/10.3389/fpsyg.2019.00463

Song, D. (2016). Student-generated questioning and quality questions: A literature review. *Research Journal of Educational Studies and Review*, *2*, 58–70.

Team, R. S. (2016). RStudio: Integrated development for R. In *RStudio*. Boston, MA: Inc.. http://www.rstudio.com/

Thomas, R. C., Weywadt, C. R., Anderson, J. L., B. Martinez-Papponi, B., & McDaniel, M. A. (2018). Testing encourages transfer between factual and application questions in an online learning environment. *Journal of Applied Research in Memory and Cognition*, *7*, 252–260. doi: https://doi.org/10.1016/j.jarmac.2018.03.007.

Tran, R., Rohrer, D., & Pashler, H. (2015). Retrieval practice: The lack of transfer to deductive inferences. *Psychonomic Bulletin & Review*, *22*, 135–140. https://doi.org/10.3758/s13423-014-0646-x

van Gog, T., & Sweller, J. (2015). Not new, but nearly forgotten: The testing effect decreases or even disappears as the complexity of learning materials increases. *Educational Psychology Review*, *27*, 247–264. https://doi.org/10.1007/s10648-015-9310-x

Wammes, J. D., Meade, M. E., & Fernandes, M. A. (2017). Learning terms and definitions: Drawing and the role of elaborative encoding. *Acta Psychologica*, *179*, 104–113. https://doi.org/10.1016/j.actpsy.2017.07.008

Weinstein, Y., McDermott, K. B., & Roediger, H. L. (2010). A comparison of study strategies for passages: Rereading, answering questions, and generating questions. *Journal of Experimental Psychology: Applied*, *16*, 308–316. https://doi.org/10.1037/a0020992

Wickham, H., Francois, R., Henry, L., & Müller, K. (2017). dplyr: A grammar of data manipulation. R package version 0.7.4. https://CRAN.R-project.org/package=dplyr

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

Wittrock, M. C. (1974). Learning as a generative process. *Educational Psychologist*, *11*, 87–95. https://doi.org/10.1080/00461527409529129

Wouters, P., van Nimwegen, C., van Oostendorp, H., & van der Spek, E. (2013). A meta-analysis of the cognitive and motivational effects of serious games. *Journal of Educational Psychology*, *105*, 249–265. https://doi.org/10.1037/a0031311

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Ebersbach M, Feierabend M, Nazari KBB. Comparing the effects of generating questions, testing, and restudying on students' long-term recall in university learning. *Appl Cognit Psychol*. 2020;34:724–736. https://doi.org/10.1002/acp.3639

APPENDIX A.: | Example slide and corresponding factual and transfer question (for the complete material, see OSF)

Factual question:

"Wilkening demonstrated that kindergartners already consider the speed and duration of movement when estimating the distance covered by animals. What is the constraint, assumed by Piaget to be prevalent among kindergartners, that was shown not to be exhibited by kindergartners?"

Transfer question:

If you generalize the findings of Wilkening on intuitive physics to the estimation of the volume of cylinders, which dimension(s) would preschoolers consider in their estimations?