The Testing Effect in the Lecture Hall: Does it Transfer to Content Studied but not Practiced?

Julia Glaser and Tobias Richter

University of Würzburg

Accepted for publication in the journal Teaching of Psychology (2023)

Author Note

Julia Glaser https://orcid.org/0000-0002-3534-7130

Tobias Richter https://orcid.org/0000-0002-0467-9044

We have no known conflict of interest do disclose.

The data files and analysis syntax underlying the analyses reported in this study and an online supplement with additional results are publicly available via Open Science Framework (https://osf.io/wc3kh/?view_only=4e9528d3ad684c36a2e8a18872a781db). Materials are available from the authors upon request.

This research was funded by the Federal Ministry of Education and Research (BMBF) in the project CoTeach (Grant no. 01JA2020). Tobias Richter's work on this article was also supported by a grant from the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) for the Research Unit "Lasting Learning: Cognitive mechanisms and effective instructional implementation" (Grant FOR 5254/1, project number 450142163). We thank Ann-Sophie Schriewer for help in data scoring. We furthermore thank Johanna Grimm, Wolfgang Lenhard, Peter Marx, Lisa Pilotek, and Sandra Schmiedeler for letting us conduct our study in their courses.

Correspondence concerning this article should be addressed to Julia Glaser or Tobias Richter, University of Würzburg, Department of Psychology IV, Wittelsbacherplatz 1, 97074 Würzburg, Germany. Email: julia.glaser@uni-wuerzburg.de or tobias.richter@uni-wuerzburg.de

Abstract

Background: Practice tests have been shown to be an effective means to foster long-term retention in higher education, at least compared to restudying (i.e., the testing effect).

Objective: The present study replicated and extended prior research by examining whether and to what extent the positive effects of testing on long-term retention in a typical psychology lecture transfers to content presented only during initial learning (and not practiced).

Method: Using a within-subjects design, we alternated post-lecture multiple-choice practice tests and restudying opportunities in two psychology classes ($N = 67$). One week after the final lecture session of a cycle of six weekly lecture sessions, retention of learning content was assessed by comparing performance on questions referring to content practiced via testing, encountered via restudying, or unreviewed.

Results: We found a testing effect for practiced content, whereas no transfer effect occurred for untested content from the same lecture sessions.

Conclusion: These results show that the testing effect is a powerful learning tool, but also suggest a possible boundary condition pertaining only to explicitly tested content.

Teaching Implications: Practice testing should be integrated regularly in higher-education courses to foster long-term retention for a final test. However, educators should take care that important content is fully covered in practice tests.

Keywords: testing effect, retrieval practice, desirable difficulties, transfer, university teaching

The Testing Effect in the Lecture Hall:

Does it Transfer to Content Studied but not Practiced?

One goal of formal education from kindergarten to higher education is to foster the acquisition of lasting knowledge (Richter et al., 2022). Test-enhanced learning, consisting of an initial study phase followed by practice tests administered in class or via self-testing, has been shown to be an effective and efficient way to foster long-term retention more broadly (Roelle et al., 2022; Rowland, 2014; Yang et al., 2021) and more specifically in psychology courses (see Schwieren et al., 2017, for a meta-analysis). In a typical implementation of retrieval practice in a university classroom, students attend a class and are provided with a practice test afterwards with or without feedback. Researchers would then compare studying plus retrieval to studying plus restudying on a retention test administered sometime after learning, usually after one week or even later (Roediger & Butler, 2011).

In the research partially replicated and extended in the current study, we asked introductory psychology students to complete short-answer questions for half of the content in the lecture session with corrective feedback (practice test) and summarizing statements (restudy) for the other half of the content; this took place during review sessions administered online in the week after the lecture session (Glaser & Richter, 2023a). Across six different lecture topics, a positive testing effect emerged in a final test one week after learning. Content for which short-answer questions were posed was remembered better than content for which summarizing statements were presented, with a medium effect size (Cohen's $d = 0.55$) (Glaser & Richter).

Meta-analytic reviews suggest that positive testing effects such as the one obtained in our recent study (Glaser & Richter, 2023a) are a very robust phenomenon. Practice testing seems to be more effective than restudying in different age groups, with different types of learning material, and in the laboratory, with an overall medium effect size (Hedge's $g = 0.50$, Rowland, 2014) and classroom settings ($g = 0.33$; Yang et al., 2021; see also Agarwal et al.,

2021). The meta-analyses and additional research also point to relevant moderators. Practice testing seems to be more effective for longer retention intervals, stressing its usefulness for fostering lasting learning (Rowland, 2014). Corrective feedback for responses in the practice test can increase the positive effect of testing by strengthening existing knowledge and allowing learners to detect knowledge gaps and to improve metacognitive calibration, which benefits future learning activities (*indirect testing effect*: Arnold & McDermott, 2013). If no feedback is given, the retrievability of the learned content is crucial because successful retrieval is a precondition for retrieval practice to be effective (*direct testing effect*; Greving et al., 2022; Greving & Richter, 2018). The questions in the practice test must stimulate active retrieval, which can be achieved through cued-recall (short-answer) questions or appropriately designed multiple-choice questions (Butler, 2018). Finally, repeated testing has been shown to be particularly beneficial for learning (Butler, 2010).

Given the robust learning effects of test-enhanced learning, practice testing is considered a silver bullet for promoting lasting knowledge (Dunlosky et al., 2013). However, the evidence for the applicability of test-enhanced learning in educational contexts is somewhat limited because in most experimental studies, the criterial test covers exactly the practiced or reviewed content after learning. Although demonstrating the effectiveness of practice tests for the explicitly tested content is important, the applicability of test-enhanced learning would be greatly expanded if practice tests would also benefit the retention of other content encountered during the initial learning session, even if the information was not explicitly tested. For instance, psychology students attending a lecture are typically expected to acquire broad knowledge of the topics covered in the lecture, and psychology educators using low-stakes quizzes in their classroom usually do so in the hope to make all the important content stick, not just the subset of facts that are covered in the practice test.

A meta-analysis by Pan and Rickard (2018) that included 192 effect sizes from studies with different types of learning materials (often expository texts) revealed a medium-sized

positive effect of practice testing across different types of transfer, including application and inference questions, transfer from one question format to another, and the transfer to content encountered during initial learning but not covered by the practice test. However, for the latter type of transfer, the research is especially inconclusive. The present field experiment replicated our previous research (Glaser & Richter, 2023a) and extended it to provide a clarification of this question in a typical lecture setting. To ground our research questions, we briefly review theoretical explanations of the testing effect that suggest practice testing might extend to content learned but not explicitly tested. We will then discuss the available empirical studies that have addressed this phenomenon in more detail.

Transfer of Test-enhanced Learning to Content Studied but not Practiced

Transfer occurs when "learning in one context or with one set of materials impacts on performance in another context or with other related materials" (Perkins & Salomon, 1992, p. 425). In the present study, we examined a specific form of transfer effects that is potentially elicited by test-enhanced learning. Specifically, we were interested in whether the beneficial effects of retrieval practice extend to content learned in the same lecture session but not explicitly covered in the practice test. Theoretical accounts of the testing effect suggest that such transfer effects are possible. For example, Carpenter (2009) proposed elaborative retrieval as a mechanism that underlies direct testing effects. According to her account, (successful) retrieval not only involves the activation of the retrieved information but also the activation of related concepts through a spreading activation mechanism. Citing an example from Anderson (1976), Carpenter illustrates the mechanism of elaborative retrieval by the attempt to learn the association between the words *dog* and *chair* by thinking of a dog who loved to sit on his master's chair but was scolded by the master for leaving his hairs on the chair. This brief story creates an elaborative retrieval structure that provides multiple pathways for future retrieval attempts (e.g., *dog-scold-chair*, *dog-master-chair*, *dog-sit-chair*, *dog-master-scold-chair*, ...), thus facilitating long-term retention. Once formed, the

elaborative structure should also facilitate retrieval of other elements of the structure such as the concept *chair* when cued with *master* or *master* when cued with *chair*.

In line with this assumption, Carpenter (2011) found that retrieval practice with word pairs not only improved cued recall of the practiced target words but also recall of words semantically associated with the word pairs. These semantic associates act as mediators and can enhance the recall of the learned materials (see also Carpenter & Yeung, 2017; Pyc & Rawson, 2010) but in this process, the retrievability of the mediators is also enhanced. In educational settings, the concepts taught within a lesson on a specific topic are usually coherent with each other, which is a favorable condition for the formation of elaborative structures. Therefore, if practice testing elicits elaborative retrieval, the memory traces of information contained in the lesson but not tested in the practice test might also be strengthened if this information is associated with the practiced content. In addition to elaborative retrieval, mechanisms assumed to underly indirect testing effects might also contribute to this specific type of transfer effect, which has also been called *retrieval-induced facilitation* (Chan et al., 2006).

In their meta-analysis of transfer effects, however, Pan and Rickard (2018) found no significant overall transfer effects for studies that examined transfer effects to the retention of untested materials seen during initial study ($d = 0.16$, $p = .20$, $k = 17$). In other words, the hypothesis of retrieval-induced facilitation received no support. However, the analysis also revealed a considerable heterogeneity of effects that warrants a closer look at the specifics of the primary studies.

Several studies included in Pan and Rickard's (2018) meta-analysis provide support of transfer effects to untested content, most of them based on expository texts (Butler, 2010, Experiments 1b, 2, and 3; Chan et al., 2006; Chan, 2010; Hinze et al., 2013, Experiment 2). In these studies, great care was taken that the transfer questions were indeed semantically highly coherent with the practiced content, which might be a critical condition for transfer effects to

occur. Other studies used specific tasks in the practice tests that might facilitate transfer effects to untested contents. For example, Hinze et al. (2013, Experiment 3) found that transfer effects occurred only if the practice test included the task to write an explanation, which is likely to foster elaborative (or constructive) retrieval, but not if the task was free recall. Another example is a classroom experiment conducted by Balch (1998) in an introductory psychology course. In this study, the students who took the practice tests additionally scored the practice test of another participant before they received feedback on their own answers. This additional task might have enhanced elaborative retrieval, thus contributing to a transfer effect. Finally, McDaniel et al. (2012) reported transfer effects in two field experiments conducted with undergraduates as part of a web-based class on biopsychological topics, one experiment (based on a very small sample) providing support, a second experiment (based on a slightly larger sample) failing to provide support for transfer effects of testing compared to restudying. Thus, the results by McDaniel et al. (2012) are inconclusive regarding the focal question.

Other studies included in the meta-analysis by Pan and Rickard (2018) also provided evidence against transfer effects of testing, at least against transfer to contents not strongly related to the practice questions. For example, in three experiments reported by Butler (2010), performance in control questions that referred to the expository text read in the study phase was worse in the practice test conditions compared to the restudy condition that included the possibility to reread whole passages (Experiments 1a, 1b, and 2). In still other studies, such as a study by Nungester and Duchastel (1982) with high-school students who read a history text, or classroom experiments by Wooldridge et al. (2014) and La Porte and Voss (1975) with undergraduates, transfer effects of testing on studied but untested content were not significant. However, like the experiments by McDaniel et al. (2012), several of these studies were based on small sample sizes and are likely to be underpowered, which limits the conclusions that can be drawn from their results. Another classroom experiment by Pilotti et al. (2009) is

likewise inconclusive regarding the focal question, as participants in the restudy condition (i.e., the control condition) received the practice questions with answers and were prompted to think about why the answer was correct. In other words, participants in the control condition engaged in an elaborative activity that might have enhanced transfer.

Rationale of the Present Study

The studies reviewed in the preceding section rely on a broad variety of practice tasks, use very different types of materials and practice tests, and are overall inconclusive on the question of whether practice tests improve long-term retention beyond the content that is explicitly tested. Given the practical relevance of this question for psychology educators, the current study addressed transfer effects of practice testing to untested contents in typical psychology courses, covering several weeks of a semester and with a broad range of topics, to explore the generalizability of results. As a replication and extension of our prior work (Glaser & Richter, 2023a), we examined whether short practice tests that supplemented six regular lecture sessions in two different psychology lectures during one semester would benefit long-term retention, measured after a retention interval of one to several weeks post-learning (one week after the review session following the last of six consecutive weekly lecture sessions). In addition to examining testing effects for the knowledge targeted by the practice questions, we focused on the question whether the positive effects of practice tests on learning would extend to other content from the same lecture, which was neither restudied nor explicitly tested. In terms of direct effects of testing, such an effect might occur when practice testing encourages elaborative retrieval, which includes the activation of content linked to the content directly targeted by the question (Chan et al., 2006).

We compared the effects of practice testing with multiple-choice questions to the effects of restudying summary statements that were equivalent to the correct responses in the practice tests (within-subjects, alternating weekly). Both the testing condition and the restudying condition were offered as online learning materials and implemented in a highly

economical way, taking only a few minutes to complete. The questions in the practice tests were presented with corrective feedback and repeated once if participants failed to provide the correct answer in the first run, which should create favorable conditions for testing effects to occur. In the final test presented one week after the final review session, participants received either multiple-choice or short-answer questions.

In this setting, we tested the following hypotheses. First, we expected a testing effect. Content from lecture sessions accompanied by practice tests should be remembered better than content from lecture sessions accompanied by restudy materials (Hypothesis 1). However, this testing effect should be stronger for questions that are directly tested in the practice tests compared to questions that refer to content from the same lecture sessions, but which are not also part of the practice tests or the restudy material (Hypothesis 2). Third, we expected learning outcomes to be better for questions that directly refer to reviewed (practiced or restudied) content than learning outcomes assessed with new questions that refer to content that was not reviewed (Hypothesis 3).

Additionally, we examined the following exploratory research questions. First, we examined whether the testing effect would be stronger for learners who performed better in the practice tests, as reflected in a higher average retrievability in the practice tests. This research question is based on theory and research showing that retrievability is crucial for retrieval practice to be effective, at least when no feedback is presented (for evidence in a lecture context, see Greving & Richter, 2018). Moreover, we explored whether the question format in the final test would make a difference in the testing effect. Insofar as transfer-appropriate processing plays a role for the testing effect (Veltre et al., 2015), beneficial effects of retrieval practice might be larger if the same question format (multiple-choice) is used in the practice and the final test compared to different question formats (multiple-choice vs. short-answer questions). Finally, by employing two types of questions in the final test, we were able to investigate whether the testing effect varies as a function of whether the learning

outcomes are assessed in the same or a different response format as in the practice test. The principle of transfer-appropriate processing (Morris et al., 1977) suggests that the congruence of response format might be important. Nevertheless, testing effects with incongruent response formats in practice and final tests have been found in previous studies (yielding a mean effect size of $d = 0.28$ in the meta-analysis by Pan & Rickard, 2018), even though they seem to be smaller than testing effects found with congruent response formats ($d = 0.58$).

Method

Participants

We conducted an online study with students enrolled in the teacher-training program at the University of Würzburg. Of the 97 students who signed up for the study, 67 completed all parts of the study. The remaining 30 students had to be excluded from the analysis because they did not participate in all parts of the study or skipped large parts of the practice tests (more than 50%). Participants whose data were excluded did not differ significantly from those whose data were included in relevant learner characteristics such as gender, age, study performance and prior knowledge (see Table S1 in the online supplement for more details; Glaser & Richter, 2023b).

Participants were recruited from two different courses (*Behavioral and Learning Disorders in Childhood and Youth* and *Developmental Psychology in Childhood and Youth*), both of which are part of the mandatory psychology curriculum for prospective teachers in their first year of study. Twenty-five students were studying to become elementary school teachers (Grundschule), four were non-academic track middle-school teachers (Mittelschule), three were studying for secondary school (Realschule), 21 for high school (Gymnasium), 11 for special education (Sonder-/Förderschule), and two were studying pedagogy (missing data from 1 student).

Most of the 67 participants included in the analysis identified as female (86.6%) and their age ranged from 18 to 37 years ($M = 20.63$, $SD = 2.92$). Only one participant reported a

native language other than German. All participants received course credits for their participation in the study. Detailed information on the sample can be found in Table S2 in the online supplement (Glaser & Richter, 2023b).

Kommentiert [JM1]: This was redundant with the above sentence. I combined the info above and omitted this one.

Kommentiert [JM2]: As the 2nd table mentioned, call this Table S2 (then S2 below becomes S3).

A sensitivity analysis revealed that assuming a power (1-β) of .90, a Type I error probability $\alpha = .05$ and a correlation $\rho = .561$ between the levels of the repeated-measures factor (the median of the observed correlations), the design was sensitive enough to detect a small testing effect of $f = 0.154$ (corresponding to $\eta^2 = .023$ or Cohen's $d = 0.308$; power analysis performed with GPower, Faul et al., 2007; effect size transformations computed with the tool provided by Lenhard & Lenhard, 2016).

Materials

Practice Tests and Summary Statements

Within each of the two courses, there were 12 sessions, for each of which 15 information units were identified. For every information unit, we constructed a summarizing statement and a short-answer or a multiple-choice question. Summarizing statements were created by summarizing the main ideas of the information unit in one short sentence (e.g., *Dyslexia is stable during development, but therapy can have a positive influence*). Short-answer questions were created by asking for the main ideas out of the information unit (e.g., *What can be said about the stability of the development of dyslexia, especially in relation to therapy? Give your answer in 1-2 short sentences*). Multiple-choice questions were created by adding four possible answers to the short-answer question with varying numbers of correct alternatives (e.g., response options: *The development of dyslexia ... (a) is stable and cannot be influenced by therapy; (b) is stable, but therapy can have a positive influence; (c) is unstable depending on the child's development in other areas; (d) is unstable, and the problem disappears in some cases without therapy*). The questions including the sample solution for all topics can be viewed in OSF in German as well as in an English translation.

For each lecture session, 10 of the 15 information units were selected and presented either as a summary statement or as a multiple-choice question. The statements were presented, and participants had up to 60 seconds to restudy them but could move on to the next item earlier. For the multiple-choice questions, participants also had up to 60 seconds to provide their responses. After their answer, they were shown corrective feedback (*correct/incorrect*) and the correct answers. If participants chose a wrong answer or missed a correct option for a question, they received the question again in the end. As before, they received feedback on their answers.

Kommentiert [JM3]: In OSF, I recommend including for each idea unit, the practice test questions and summary statement, and criterial test questions. Then direct readers there, noting that the materials are available in German (if you choose not to translate these).

Final Test

The dependent variable was based on the performance on a final test with a total of 60 questions, 10 for each of the previously learned topics. For each topic, five questions referred to information units that had been presented or tested in the practice phase, the other five referred to information units not presented in the practice phase.

The final tests were either presented with multiple-choice questions (n = 30) or with short-answer questions (n = 37). The distribution of participants among question format was randomized and balanced within the two lectures (see Table S3 in the online supplement, Glaser & Richter, 2023b).

The multiple-choice questions were scored by partial credit, with 0.25 points given for each response option that was correctly ticked or correctly not ticked. The short-answer questions were also scored according to a partial credit scheme (0-1 points in intervals of 0.25). Two raters independently coded the responses to 120 questions from 12 participants (60 questions for each participant) to estimate inter-rater reliability of the short-answer questions. Cohen's κ was .816 (SE = .017), indicating almost perfect agreement according to Landis and Koch (1977). The interrater agreement for each topic is provided in Table S4 in the online supplement (Glaser & Richter, 2023b). Given the high inter-rater reliability, the remaining answers to the short-answer questions were coded by only one coder.

Retrievability

For each participant, a mean retrievability score was computed based on the sum of correctly answered questions in the practice tests. The minimum score for each question was 0 (none of the four response options correctly ticked or left unticked), the maximum score was 1 (all four response options correctly ticked or left unticked). Each correctly ticked or not ticked option was scored with 0.25 points.

Additional Measures

For exploratory purposes, we also assessed need for cognition with the short form of the Need for Cognition Scale (Bless et al., 1994). The exploratory analyses with this scale yielded no interesting insights. Therefore, the results will not be reported. Participants also provided single-item ratings of their prior knowledge for the six topics covered in the chosen lecture (used for the description of the sample) and the comprehensibility and usefulness of the study for their own learning activities on 5-point scales. Participants also indicated whether they had used other learning methods than those implemented in the review sessions. Results concerning these measures are provided in the online supplement (Glaser & Richter, 2023b).

Design

We investigated the effect of learning condition (restudy vs. testing) on the final test performance (either questions referring to restudied or tested content vs. new content). Learning type and type of test question were varied within-subjects. Each participant received practice tests for three out of six lecture topics and summarizing statements to restudy three other lecture topics. The assignment of lecture topics to either the testing or the restudy condition as well as the sequence of conditions was counterbalanced between participants. Each participant was randomly assigned to one of the two sequences. The final test, which covered all previous lecture topics, was presented as either short-answer or multiple-choice questions and included questions that had previously been asked in the same wording as well

as questions that had previously been presented as a statement and questions that addressed related, untested knowledge.

In sum, the study was based on a 2 (learning condition: testing vs. restudy) X 2 (question type in the final test: restudied/tested vs. new content) X 2 (question format in the final test: short-answer vs. multiple-choice tests) within-subjects design. Moreover, the lecture varied between participants as a quasi-experimental factor with two levels (*Behavioral and Learning Disorders in Childhood and Youth* vs. *Development in Childhood and Youth*).

Procedure

Data for the study were collected from November 2021 until January 2022 with the online survey tool SoSci Survey (www.soscisurvey.de). Across all topics, participants estimated their prior knowledge on the lecture topic as moderate ($M = 2.47$, $SD = 0.70$, on a 5-point scale from 1 to 5). The distribution of participants between lecture topics and means and standard deviations for their self-rated prior knowledge for each topic are provided in Table S3 (Glaser & Richter, 2023b).

After choosing the course they attended (*Behavioral and Learning Disorders in Childhood and Youth* or *Development in Childhood and Youth*), participants provided their demographic information and received the *Need for Cognition* scale, and the self-assessment of prior knowledge. They also provided their email address to register for the subsequent parts of the study.

During the course, students attended their classes as usual. Two days after the first lecture lesson, which was part of the study (*Reading and Writing Problems* or *Development of Thinking*), participants received their first link with the review task via email. They were asked to respond to the practice questions or restudy the summarizing statements to review the core contents of the lecture session. Participants were randomly assigned to either the restudy condition or the testing condition. One week later, two days after the second thematic lesson, they received the second link with a new review task via email. Those participants

assigned to the testing condition in the first review session now received the summarizing statements, whereas those participants assigned to the restudy condition in the first review session now received the practice test. This procedure was repeated for Sessions 3 and 4 and again for Sessions 5 and 6. Therefore, each participant received the restudy condition and the practice session each three times, in an alternating sequence.

The final part of the study took place one week after the completion of the final review session. Participants now received the final test with 60 questions (either multiple-choice or short-answer questions), covering all six themes of the previous lessons and containing questions referring to practiced/restudied content and questions referring to lecture content that was not also part of the review session. Finally, participants rated the comprehensibility and usefulness of the study for their own learning activities and indicated whether they had used other learning methods than those implemented in the review sessions. Figure 1 illustrates the design and procedure of the experiment.

+++ Figure 1 about here +++

Results

We estimated a linear model (mixed ANOVA) with learning condition and question type as experimental factors varied within-subjects. Moreover, type of final test and lecture were included as between-subject factors. The model also included the interactions of these variables. We used an α-level of .05 for all statistical tests. We report partial η^2 as the effect size measure. All data files and analysis scripts can be found in the online repository (Glaser & Richter, 2023b).

Effects of Learning Condition, Question Type, and Question Format

We found no main effect for learning condition, $F(1, 64) = 0.63$, $p = .429$, $\eta_p^2 = .01$. Thus, Hypothesis 1, predicting an overall testing effect, could not be supported. Importantly, however, the interaction of learning condition and question type was significant, $F(1, 64) = 5.18$, $p = .026$, $\eta_p^2 = .08$. Figure 2 displays the interaction of learning condition and question

type. Follow-up tests revealed a testing effect for questions referring to reviewed content. Final test performance in questions that referred to tested contents (M = .70, SE = .020) was better than performance in questions that referred to restudied contents (M = .66, SE = .017), $F(1, 64) = 5.33$, $p = .024$, $\eta_p^2 = .08$. In contrast, for questions referring to lecture content that was not reviewed, we found no significant difference between sessions followed by testing (M = .62, SE = .018) compared to restudying (M = .63, SE = .017) , $F(1, 64) = 0.74$, $p = .393$, η_p^2 = .01. Thus, Hypothesis 2 that testing would especially benefit the tested content was supported. More specifically, a testing effect occurred only for content explicitly tested and no transfer effect to untested content from the same lecture occurred.

Kommentiert [JM4]: I think putting this part up front in the sentence helps readers to follow along.

+++ Figure 2 about here+++

We found a strong main effect for question type, with overall better learning outcomes in the final test for questions referring to reviewed (i.e., tested or restudied) content (M = .68, SE = .017) compared to questions referring to content not reviewed after the lecture (M = .63, SE = .014), $F(1, 64) = 19.09$, $p < .001$, $\eta_p^2 = .23$. Therefore, Hypothesis 3 predicting a better learning outcome for questions referring to reviewed (tested or restudied) content was basically supported. However, the interaction effect of question type and question format was also significant, $F(1, 64) = 5.83$, $p = .019$, $\eta_p^2 = .08$. Figure 3 displays the interaction of question type and question format. Follow-up tests revealed that the better learning outcome for reviewed content was significant only with multiple-choice questions (reviewed content: M = .75, SE = .024; content not reviewed: M = .67, SE = .021), $F(1, 64) = 21.19$, $p < .001$, η_p^2 = .25. In contrast, performance on short-answer questions was not significantly different between the two question types (reviewed content: M = .60, SE = .022; content not reviewed: M = .58, SE = .019), $F(1, 64) = 2.16$, $p = .147$, $\eta_p^2 = .03$.

+++ Figure 3 about here+++

We also tested whether lecture topic would moderate the hypothesized effects, which would provide information about the generalizability of the effects. Only the interaction of

lecture topic with question type was significant, $F(1, 64) = 6.67$, $p = .012$, $\eta_p^2 = .09$. Follow-up tests revealed that learning outcomes for reviewed content compared to lecture content that was not reviewed was significant only in the *Behavioral and Learning Disorders in Childhood and Youth* lecture (reviewed content: $M = .71$, $SE = .021$; content not reviewed: $M = .63$, $SE = .019$), $F(1, 64) = 28.60$, $p < .001$, $\eta_p^2 = .31$, but not in the *Development in Childhood and Youth* lecture, $F(1, 64) = 1.38$, $p = .244$, $\eta_p^2 = .02$ (reviewed content: $M = .64$, $SE = .025$; content not reviewed: $M = .62$, $SE = .022$). None of the other interactions of lecture topic with any of the other independent variables was significant (for all tests, $p > .103$). An overview of all effects and interactions is provided in Table S5 in the online supplement (Glaser & Richter, 2023b).

Kommentiert [JM5]: Double-check/modify numbering to make sure they follow along with order of presentation in the manuscript.

Moderating Effect of Retrievability

In an additional step, we estimated an expanded linear model to explore whether the testing effect found in the primary analysis would be moderated by the average retrievability, operationalized as performance in the practice tests. When adding retrievability to the model, we found a significant main effect on final test performance, $F(1, 63) = 10.81$, $p = .002$, $\eta_p^2 = .15$, with the average higher retrievability being associated with better learning outcomes ($r = .40$). However, none of the interactions with learning condition was significant (for all tests of two- and three-way interactions), and all the hypothesis-relevant effects reported above remained basically unchanged after including retrievability in the model (see Table S6 in the online supplement; Glaser & Richter, 2023b).

Discussion

The present study replicated and extended our prior research (Glaser & Richter, 2023a), specifically examining whether short practice tests would benefit long-term retention compared to restudying the information and whether this effect would extend to non-reviewed content that was featured in the same lecture session. We created a minimal intervention over a period of 6 weeks in two regular university lectures, with six topics each and a within-

participant and within-topic variation of practice testing versus restudying. Learning outcomes were assessed after the review session following the final lecture in a cycle of six consecutive weekly lecture sessions, implying a retention interval of 1 to 6 weeks after learning, depending on when in the cycle the topic was learned. The main finding was that a testing effect occurred only for content that was part of the practice test but not for content that was featured in the lecture but not tested. We also found that reviewed content (i.e., tested or restudied) was overall remembered better than content that was only encountered during the lecture, at least for multiple-choice questions. Finally, we found no moderating effect of response congruency in the practice tests (containing short-answer questions) nor the final test (containing short-answer questions versus multiple-choice).

These results contribute to the literature on transfer effects of testing by providing clear evidence against the assumption of transfer effects to untested material. The significant interaction obtained in the present study demonstrates that the testing effect for questions that appeared in the practice effect and the final test is larger than the transfer effect for final test questions that refer to untested content. Moreover, our study was adequately powered, suggesting that the null effect for the transfer effect of testing was not due to a low sensitivity. Apart from these statistical arguments, a strength of the present study is its implementation in two psychology lectures based on curricular content distributed over 12 different topics. With these features, the results are informative for the application of practice testing in typical psychology courses in higher education. Our results suggest, in line with numerous other studies conducted in the university classroom (see Yang et al., 2021), that practice tests are an effective way to increase retention of the tested information. This effect did not depend on whether the question format in the practice tests (multiple-choice) matched the question format in the final test (multiple-choice versus short-answer). Our results also suggest more clearly than previous research that teachers should not expect that these positive effects extend to untested information from the same course. One practical recommendation based on

these findings is that teachers should attempt to cover the central content of their course in a practice test, to maximize the chance that the relevant knowledge lasts.

Restudied contents were also remembered better than contents not covered in the review, which indicates that restudying after a delay also has beneficial effects for long-term retention (see also Rawson & Kintsch, 2005). However, this positive effect was not as large as the effect of practice testing and occurred only in one of the two lectures and only for multiple-choice items in the final test, suggesting that teachers should use quizzing rather than restudy opportunities to support student learning in their courses.

Boundary Conditions of Transfer Effects

Do our results refute the results from previous studies that suggest a transfer effect of testing to untested information? Certainly not. A closer look at the studies providing clear positive evidence for this kind of transfer effect reveals that these studies combined the review session with specific additional tasks (explanation tasks or expectation of inference questions in the final test, Hinze et al., 2013; scoring another participant's practice test, Balch, 1998), or that they used carefully controlled materials to ensure that the untested content had indeed strong semantic associations with the content covered in the practice test (e.g., Chan et al., 2006). In our experiment, tested and untested materials were always associated by being presented within one thematically coherent unit, but the associations were otherwise likely to vary as it is typical for lecture content. Thus, our results do not rule out the possibility of transfer effects of testing to content learned but not practiced. Such effects seem to depend on certain boundary conditions, which should be systematically examined in future research. One such condition is that the strength of the semantic association of information or the coherence of the topics presented in one learning session is crucial for a transfer effect of testing to untested content, as highlighted by the elaborative processing account of the testing effect (Carpenter, 2009) or the notion of retrieval-induced facilitation (Chan et al., 2006). Another condition that could favor transfer effects is the combination of practice testing with

additional tasks or instructions that stimulate elaborative processing or generative learning (Richter et al., 2022; Roelle et al., 2022). Such measures have been shown to enhance the long-term effects of other desirable difficulties in authentic educational contexts. For example, Ziegler and Stern (2014) have shown that instructional support that enhances comparing and contrasting enhances the effectiveness of interleaving in mathematics education.

Limitations and Future Directions

Despite the clear pattern of results and the practical implications, the present study also has some limitations. One limitation already stated is that because of the implementation in the context of a regular university lecture, the strength of the semantic associations between different content featured in one lecture session could not be controlled, which limits the theoretical value of the study. Other limitations, include the lack of counterbalancing the order of topics, the varying retention interval (1-6 weeks) for each lecture topic, and the possibility that students engaged in additional learning activities during the study. Importantly, however, none of these methodological characteristics represent a confound that would undermine the internal validity of the experiment and its central conclusions.

In addition, the dropout rate in the present study was high due to the voluntary character of the study and the multiple sessions that were required for completing the experiment. Future research should find ways to keep the rate lower, for example by making study participation a mandatory part of class assignments. Finally, even though we took efforts to maximize generalizability of results, by including two different lectures, which are typical for introductory psychology lectures, and 12 different lecture topics in total, the present effects might still depend on the domain, the complexity of the topic, or didactical features of the lecture. Likewise, the population of teacher trainees at a German university that our sample was based on is relatively heterogeneous regarding their characteristics such as interests and skills, which depend, among other things, on their field of studies.

Nevertheless, whether and to what extent our findings generalize to other populations of learners with different backgrounds and different cognitive and motivational prerequisites is unclear. Future studies should address this issue by examining more heterogeneous samples.

Conclusion

In summary, the results of the present study underscore once more the utility that practice tests can have in university teaching for consolidating acquired knowledge. However, our findings clarify an important limitation of this benefit. The practice tests seem to selectively promote only the retention of explicitly tested content and not the retention of other content presented in the same learning unit that was not tested. To promote comprehensive learning in this context, practice tests arguably need to be supplemented by other instructional measures that promote elaborative processing and the construction of integrated mental representations of the learning content. Psychology instructors are encouraged to include all important topics in the quizzes they provide as learning opportunities for their students.

References

Agarwal, P. K., Nunes, L. D., & Blunt, J. R. (2021). Retrieval practice consistently benefits student learning: A systematic review of applied research in schools and classrooms. *Educational Psychology Review, 33*(4), 1409–1453. https://doi.org/10.1007/s10648-021-09595-9

Anderson, J. R. (1976). *Language, memory, and thought.* Lawrence Erlbaum.

Arnold, K. M., & McDermott, K. B. (2013). Test-potentiated learning: Distinguishing between direct and indirect effects of tests. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*(3), 940–945. https://doi.org/10.1037/a0029199

Balch, W. R. (1998). Practice versus review exams and final exam performance. *Teaching of Psychology, 25*(3), 181–185. https://doi.org/10.1207/s15328023top2503_3

Bless, H., Wänke, M., Bohner, G., Fellhauer, R. F., & Schwarz, N. (1994). Need for Cognition: eine Skala zur Erfassung von Engagement und Freude bei Denkaufgaben [Need for Cognition: A scale measuring engagement and enjoyment in cognitive tasks]. *Zeitschrift für Sozialpsychologie*, *25*, 147-154.

Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(5), 1118–1133. https://doi.org/10.1037/a0019902

Butler, A. C. (2018). Multiple-choice testing in education: Are the best practices for assessment also good for learning? *Journal of Applied Research in Memory and Cognition, 7*(3), 323-331. https://doi.org/10.1016/j.jarmac.2018.07.002

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

Carpenter, S. K., & Yeung, K. L. (2017). The role of mediator strength in learning from retrieval. *Journal of Memory and Language, 92*, 128-141. https://doi.org/10.1016/j.jml.2016.06.008

Chan, J. C. (2010). Long-term effects of testing on the recall of nontested materials. *Memory, 18*(1), 49-57. https://doi.org/10.1080/09658210903405737

Chan, J. C. K., McDermott, K. B., & Roediger, H. L. III. (2006). Retrieval-induced facilitation: Initially nontested material can benefit from prior testing of related material. *Journal of Experimental Psychology: General, 135*(4), 553–571. https://doi.org/10.1037/0096-3445.135.4.553

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

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175-191. https://doi.org/10.3758/BF03193146

Glaser, J., & Richter, T. (2023a). The testing effect in the lecture hall: Does it depend on learner prerequisites? *Psychology Learning & Teaching, 22*(2), 159–178. https://doi.org/10.1177/14757257221136660

Glaser, J., & Richter, T. (2023b, October 24). *The testing effect in the lecture hall: Transfer to Untested Content?* [Supplemental online material]. https://osf.io/wc3kh/?view_only=23bc6a041d4a467b87788fbe2cc3bd9b

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

Greving, S., Lenhard, W., & Richter, T. (2023). The testing effect in university teaching: Using multiple-choice testing to promote retention of highly retrievable information. *Teaching of Psychology, 50*(4), 332–341. https://doi.org/10.1177/00986283211061204

Hinze, S. R., Wiley, J., & Pellegrino, J. W. (2013). The importance of constructive comprehension processes in learning from tests. *Journal of Memory and Language, 69*(2), 151–164. https://doi.org/10.1016/j.jml.2013.03.002

LaPorte, R. E., & Voss, J. F. (1975). Retention of prose materials as a function of postacquisition testing. *Journal of Educational Psychology, 67*(2), 259–266. https://doi.org/10.1037/h0076933

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310

Lenhard, W., & Lenhard, A. (2016). *Computation of effect sizes* [Online tool]. Psychometrica. https://www.psychometrica.de/effect_size.html

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

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior, 16*(5), 519–533. https://doi.org/10.1016/S0022-5371(77)80016-9

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

Pan, S. C., & Rickard, T. C. (2018). Transfer of test-enhanced learning: Meta-analytic review and synthesis. *Psychological Bulletin, 144*(7), 710–756. https://doi.org/10.1037/bul0000151

Perkins, D. N., & Salomon, G. (1992). Transfer of learning. In T. Husén & T. N. Postlethwaite (Eds.), *The international encyclopedia of education* (2nd ed., pp. 425-441). Pergamon.

Pilotti, M., Chodorow, M., & Petrov, R. (2009). The usefulness of retrieval practice and review-only practice for answering conceptually related test questions. *The Journal of General Psychology, 136*(2), 179-204. https://doi.org/10.3200/GENP.136.2.179-204

Pyc, M. A., & Rawson, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. *Science, 330*(6002), 335. https://doi.org/10.1126/science.1191465

Rawson, K. A., & Kintsch, W. (2005). Rereading effects depend on time of test. *Journal of Educational Psychology, 97*(1), 70-80. https://doi.org/10.1037/0022-0663.97.1.70

Richter, T., Berger, R., Ebersbach, M., Eitel, A., Endres, T., Borromeo Ferri, R., Hänze, M., Lachner, A., Leutner, D., Lipowsky, F., Nemeth, L., Renkl, A., Roelle, J., Rummer R., Scheiter, K., Schweppe J., von Aufschnaiter, C., & Vorholzer, A. (2022). How to promote lasting learning in schools: Theoretical approaches and an agenda for research. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie/German Journal of Developmental Psychology and Educational Psychology, 54*(4), 135-141. https://doi.org/10.1026/0049-8637/a000258

Roediger, H. L. III, & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences, 15*(1), 20–27. https://doi.org/10.1016/j.tics.2010.09.003

Roelle, J., Schweppe, J., Endres, T., Lachner, A., von Aufschnaiter, C., Renkl, A., Eitel, A., Leutner, D., Rummer, R., Scheiter, K., & Vorholzer, A. (2022). Combining retrieval practice and generative learning in educational contexts: Promises and challenges. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie/German Journal of Developmental Psychology and Educational Psychology, 54*(4), 142–150. https://doi.org/10.1026/0049-8637/a000261

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

Schwieren, J., Barenberg, J., & Dutke, S. (2017). The testing effect in the psychology classroom: A meta-analytic perspective. *Psychology Learning & Teaching, 16*(2), 179-196. https://doi.org/10.1177/1475725717695149

Veltre, M. T., Cho, K. W., & Neely, J. H. (2015). Transfer-appropriate processing in the testing effect. *Memory, 23*(8), 1229-1237. https://doi.org/10.1080/09658211.2014.970196

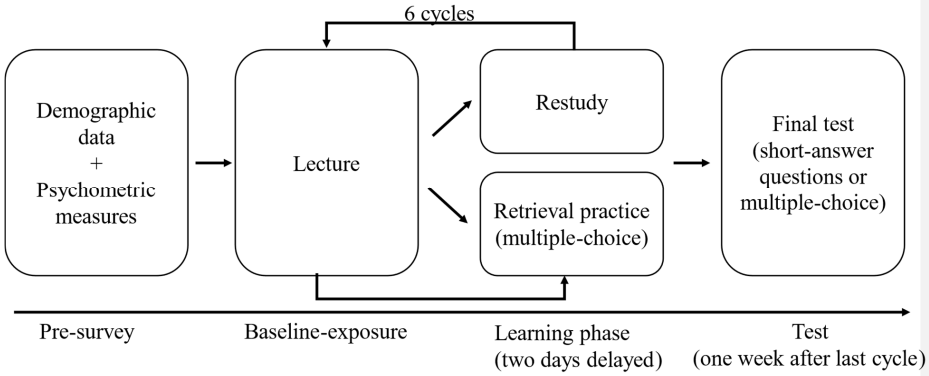
Wooldridge, C. L., Bugg, J. M., McDaniel, M. A., & Liu, Y. (2014). The testing effect with authentic educational materials: A cautionary note. *Journal of Applied Research in Memory and Cognition, 3*(3), 214–221. https://doi.org/10.1037/h0101801

Yang, C., Luo, L., Vadillo, M. A., Yu, R., & Shanks, D. R. (2021). Testing (quizzing) boosts classroom learning: A systematic and meta-analytic review. *Psychological Bulletin*, *147*(4), 399–435. https://doi.org/10.1037/bul0000309

Ziegler, E., & Stern, E. (2014). Delayed benefits of learning elementary algebraic transformations through contrasted comparisons. *Learning and Instruction, 33*, 131-146. https://doi.org/10.1016/j.learninstruc.2014.04.006

Figure 1

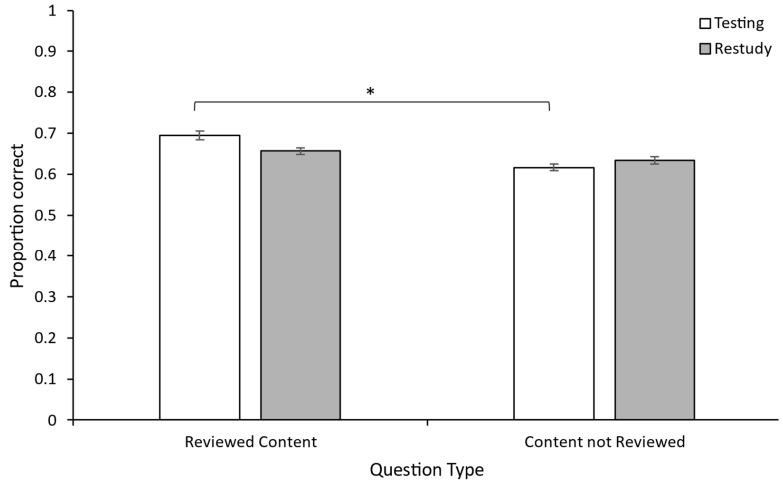
Flow Diagram Illustrating the Design and Procedure of the Study



Note. Participants from two different lectures could volunteer to participate in the experiment. In each lecture, demographic data and psychometric measures were collected first, followed by the baseline exposure to one of the lecture sessions. Two days after the session, the review phase took part with either retrieval practice or restudy (within-subjects, alternating weekly). These two steps were repeated weekly for 6 weeks. One week after the final lecture session, a test was administered, which covered all former lecture topics. The test was either presented with short-answer or multiple-choice questions.

Figure 2

Interaction of Learning Condition and Question Type

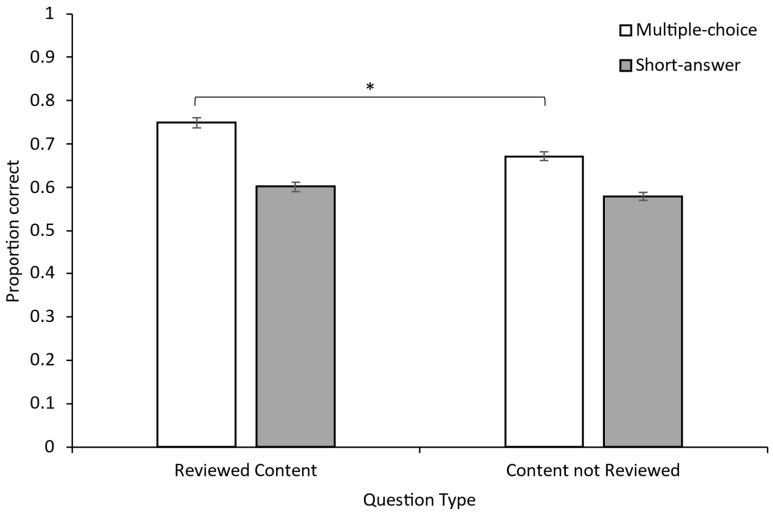


Note. Error bars represent the standard error of the mean. The *y*-axis shows the proportion of correct responses in the final test. The interaction of learning condition and question type was significant, $p = .026$.

Figure 3

Interaction of Question Type and Question Format

Kommentiert [JM6]: Can you add asterisks here to denote the significant comparison between question types?



Note. Error bars represent the standard error of the mean. The *y*-axis shows the proportion of correct responses in the final test.