Received: 6 July 2021 | Revised: 11 February 2022 | Accepted: 30 March 2022

DOI: 10.1002/acp.3943

WILEY

SHORT PAPER

Are open-book tests still as effective as closed-book tests even after a delay of 2 weeks?

Kristin Wenzel[1] | Judith Schweppe[2] | Ralf Rummer[1]

[1]Department of Psychology, University of Kassel, Kassel, Germany

[2]University of Passau, Passau, Germany

Correspondence
Ralf Rummer, Department of Psychology, University of Kassel Holländische Straße 36-38, 34127 Kassel, Germany.
Email: rummer@uni-kassel.de

Abstract

The present work was conducted to re-examine the findings of Agarwal et al. (*Applied Cognitive Psychology*, 22(7), 861–876, 2008), which showed that both closed-book tests (with feedback) and open-book tests increased learning outcomes after 1 week compared to simple re-study of the same materials. However, contrary to often found benefits of retrieval practice—which should be more pronounced in closed-book tests—both test conditions proved to be similarly effective. As retrieval practice benefits increase with retention interval, this pattern may change with a longer delay. Hence, we conducted a laboratory study and applied three within-participant learning conditions (re-study, open-book test, closed-book test with feedback) with a 2 weeks instead of 1 week delay between studying and the final test. Notably, our results mirrored the findings of Agarwal et al. (*Applied Cognitive Psychology*, 22(7), 861–876, 2008) showing that open-book and closed-book tests outperform re-study but are similarly effective—even using a slightly changed procedure, new materials, a different sample, and a longer delay.

KEYWORDS
closed-book tests, delayed final test performance, open-book tests, retrieval practice, testing effect

1 | INTRODUCTION

Recent work has repeatedly shown that taking practice tests or quizzes on previously studied information increases learners' long-term learning compared to re-reading or note-taking. This beneficial effect is known as *retrieval practice effect*, *testing effect*, *or test-enhanced learning* (e.g., Adesope et al., 2017; Rowland, 2014; Yang et al., 2021). Such benefits of tests arise in laboratory settings and in naturalistic applied learning contexts like school or university classes, for varying (curricular) materials, and when using different forms of test questions (e.g., Adesope et al., 2017; Agarwal et al., 2021; Batsell et al., 2017; Dunlosky et al., 2013; Karpicke & Aue, 2015; Rowland, 2014). Interestingly, learners do not seem to be aware of these positive effects of tests and often expect to profit more from re-studying than from test-taking, at least as reflected in judgments of learning (e.g., Karpicke & Blunt, 2011; Roediger & Karpicke, 2006; but see also Weissgerber & Rummer, 2022).

The benefits of tests are often explained by increased retrieval practice that elicits deeper and more elaborate processing as well as better anchoring of the information in long-term memory (e.g., Bjork & Bjork, 1992, 2011; Carpenter, 2009; Dunlosky et al., 2013; Rowland, 2014). The positive effects are also attributed to the higher difficulty of the retrieval task and the increased effort that is needed to retrieve the information (e.g., Pyc & Rawson, 2009; Rowland, 2014; see also *desirable difficulties*, R. A. Bjork, 1994). Additionally, higher retrieval success was often linked to higher long-term learning (especially when no feedback was given after the tests; e.g., Pyc & Rawson, 2009; Richland et al., 2005; Rowland, 2014).

Importantly, there are several further moderators of the effectiveness of tests. This includes the final test delay, that is, the time interval between the end of the learning phase and the final test assessing learning outcomes (e.g., Rowland, 2014): It has been demonstrated that benefits of tests over simple and passive re-studying are stronger with longer compared to shorter delays. Notably, there is also evidence indicating that tests outperform stronger control conditions (like note-taking) only after a longer delay: For instance, Rummer et al. (2017) contrasted tests, re-reading, and note-taking in retention intervals of 5 min, 1 week, and 2 weeks. Note-taking outperformed testing and re-reading in the 5 min condition; in the 1-week condition, both note-taking and testing were more effective than re-reading, but did not differ from each other. With a final test delay of 2 weeks, however, testing outperformed both re-reading and note-taking. Based on these findings, it is plausible to assume that retrieval practice—at least compared to more elaborate control conditions—is particularly effective for retention intervals longer than 1 week.

The typically applied tests can be subsumed as *closed-book tests* because learners are not allowed to consult the previously studied materials while answering the test questions or while responding to the recall cue but have to retrieve the information from memory. Hence, learners must engage in retrieval practice that necessitates effort to correctly retrieve the information and to successfully overcome the posed difficulty, which in turn strongly benefits their long-term learning (e.g., R. A. Bjork, 1994; Pyc & Rawson, 2009; Rowland, 2014). However, it was also argued that *open-book tests*—in which learners are allowed to consult the previously studied materials while answering the test questions—might be at least as effective as closed-book tests: Open-book tests are supposed to facilitate higher level thinking skills or elaborate processing, to lead to more accurate mental representations of the learning content, and to elicit more correctly answered test questions—all without affording difficult retrieval (e.g., Agarwal et al., 2008; Feller, 1994; Jacobs & Chase, 1992; Richland et al., 2005; Waldeyer et al., 2020; for reviews concerning varying aspects of open-book tests or open-book examinations see also Durning et al., 2016; Jensen & Moore, 2009). Open-book tests do not automatically exclude active retrieval of information, but it is highly plausible that learners only engage in retrieval of easier-to-retrieve information and consult the materials when answering questions regarding information that is harder to retrieve. The materials are thereby often seen as a form of immediate feedback concerning learners' practice test performance (cf. Agarwal et al., 2008). In contrast, closed-book tests require retrieval of both easy- and hard-to-retrieve information, which should in turn—at least for successfully retrieved information or when later feedback is given—lead to more durable representations and learning outcomes compared to open-book tests.

To examine the different effects of closed-book tests and open-book tests, Agarwal et al. (2008) conducted two experiments in which learners studied varying prose passages in six (Experiment 1) or eight learning conditions (Experiment 2) in a within-participant design. Among varying control conditions (e.g., no-study or multiple repeated study), they applied two open-book test conditions (study plus open-book test and open-book test with simultaneous study) and two closed-book test conditions (study plus closed-book test and study plus closed-book test with feedback). The practice tests each included seven short-answer questions based on facts and ideas covered in the respective prose passages. Notably, both experiments yielded beneficial effect of tests, insofar as learning conditions including tests resulted in higher learning outcomes on a 1 week delayed final test compared to when the materials were only restudied. Furthermore, although learners initially answered more test questions correctly in the open-book tests compared to the closed-book tests, there were no differences on the final test after studying with an open-book test or with a closed-book test with feedback. Agarwal and Roediger (2011) similarly found that although open-book tests initially resulted in higher performance, both open-book and closed-book tests yielded similar learning outcomes after a delay of 2 days. Notably, these results arose even though Agarwal and Roediger (2011) used materials that were specifically designed for open-book tests and included not only questions concerning separated facts or triggering the recall of single pieces of information but that also included comprehension and transfer questions requiring higher-order thinking like elaboration and integration of information across the respective passages. Generative tasks in closed-book or open-book styles also resulted in similar learning outcomes after 1 week in a study from Waldeyer et al. (2020). In further recent studies, after a delay of 2 days, learners profited just as much from writing closed-book essays than from writing open-book essays on two passages about astronomy (Arnold et al., 2021). Given the previously mentioned explanation of the testing effect in terms of retrieval practice, this seems surprising and may suggest that potential benefits of open-book tests and closed-book tests outweighed one another.

However, contrary findings also exist. For instance, generating questions was shown to be more beneficial for learning outcomes after 1 week when learners had access to the learning materials in an open-book style compared to when they could not consult the materials while generating (Ebersbach, 2020). Moreover, Roelle and Berthold (2017) highlighted the importance of the complexity of applied adjunct questions that were either implemented in a closed-book or an open-book style and that were provided together with expository texts on chemistry: After 1 week the net benefit of closed-book questions (compared to open-book questions) was higher for low-complexity questions than for higher complexity questions. Notably, the authors also emphasized that the benefit of implementing the adjunct questions in a closed-book style increased with an increasing delay of the final test. A recent field study further compared the long-term learning effects of open-book tests and closed-book tests in two parallel university courses concerning cognitive psychology (Rummer et al., 2019). The tests included short-answer questions focusing on central aspects of varying papers that had been covered in the respective lectures and that required a few sentences as answers. In the open-book test condition learners were thereby instructed to consult the provided learning materials and their own notes while answering the test questions. The study yielded higher long-term learning for

learners in a closed-book test condition compared to learners in an open-book test condition—both regarding a final test during the semester as well as a module examination at the end of the semester (Rummer et al., 2019; but see also Rummer & Schweppe, 2022). Notably, the practice tests were applied at the end of seven lessons and the final test was conducted 1 week after the seventh lesson. Hence, the delay between the practice tests and the final test ranged from 1 to 7 weeks (with an additional delay of 6 weeks for the module examination). This finding might indicate that, compared to open-book tests, the benefits of effortful and difficult retrieval practice elicited by closed-book tests only arise in the long run.

In this respect, parallels might be drawn between open-book tests and note-taking: while open-book tests consist of answering test questions with the learning materials at hand, note-taking resembles a free recall task with the learning materials at hand. Also similar to note-taking, open-book tests can be regarded as a stronger control condition compared to which the benefits of testing with a higher degree of retrieval practice may only show up after longer delays. Thus, the fact that Agarwal et al. (2008) (see also, Agarwal & Roediger, 2011; Arnold et al., 2021) observed no differences between beneficial effects of open-book and closed-book tests might be attributed to the rather short final test delays of 1 week at the most. Nonetheless, given the many differences between the just described experiments, further research is necessary.

1.1 | The present research

Following this line of reasoning, we think that it is valuable to re-examine the classic work of Agarwal et al. (2008) by conceptually replicating the critical comparisons of their experiments using a longer final test delay of 2 weeks (rather than 1 week), different materials, and a slightly changed procedure (e.g., by implementing time limits). Thus, we conducted a laboratory experiment based on the experiments conducted by Agarwal et al. and included the three critical learning conditions in a within-participant design: re-study as a typical control condition, open-book tests, and closed-book tests with feedback.

As our first hypothesis, we assume that both learning conditions including tests lead to higher final test performances than re-studying. As our second hypothesis, we assume that, given the longer delay, closed-book tests lead to higher final test performances than open-book tests.

Due to the general importance of (sufficient) prior knowledge for learning and for the effectiveness of difficult and challenging tasks (e.g., Bjork & Bjork, 2011; McNamara et al., 1996)—and because previous work resulted in contrary findings concerning potentially moderating effects of prior knowledge on the effectiveness of practice tests (for a recent overview see Buchin, 2021)—we will additionally explore potential effects of self-reported prior knowledge (these further analyses and further information can be found in Appendix B).

2 | METHOD

2.1 | Participants

The size of our sample was intended to be similar and only slightly bigger than the sample sizes of the two experiments conducted by Agarwal et al. (2008). Additionally, our sample should be comparable with the samples recruited by Agarwal et al. We thus recruited a convenience sample that consisted of $N = 63$ university students. All of them were German native speakers. One participant had to be excluded because most of the data was not recorded properly. Hence, our final sample consisted of $N = 62$ participants ($M_{age} = 22.47$, $SD_{age} = 3.26$, range = 19–39; 54 females, 8 males). Before starting, participants gave informed consent. After completing both sessions of the experiment, they received course credit or cinema vouchers.

2.2 | Materials

We selected three prose passages that were each about 500 words in length ($M_{words} = 492.66$). The prose passages covered information about zen meditation (Hawkins, 2000), everyday life in Weimar of the Goethe-era (late 18th to early 19th century; Klaus, 1990), and performance and performance-motivation in old age (Bamberg et al., 2012). Where necessary, they were adapted to the current spelling (see Appendix A for descriptive statistics regarding the prose passages).

Following Agarwal et al. (2008), we constructed seven short-answer test questions for each prose passage. The test questions focused on information described in the respective passage and appeared in the order in which the information occurred in the passages (see Appendix A for an example). The same questions were later applied in the final test.

2.3 | Procedure

All instructions and countdowns for time limits of the following tasks were presented on a computer. The prose passages, all tests, and all further items were presented and worked on in a paper-pencil format. Participants were tested in small groups or individually.

The experiment consisted of two sessions that strongly mirrored the sessions and procedures of the experiments conducted by Agarwal et al. (2008). In the first session, the learning phase took place: We applied three within-participant learning conditions: a re-study condition in which participants studied the materials twice, a study and open-book test condition in which participants studied the materials once and then took a test while consulting the materials, and a study and closed-book test condition in which participants studied the materials once and then took a test (with later feedback). The three prose passages were presented in the same order for all participants, but the order in which the learning conditions were applied was counterbalanced (Version 1: re-study, closed-book test, open-book test; Version 2: open-book test, re-study, closed-book test;

Version 3: closed-book test, open-book test, re-study). In each learning condition, participants were given 5 min to initially study the respective passage (they also reported how many times they read the passages during this time: $M = 1.52$, $SD = 0.50$; $N = 61$ due to missing data). Following this initial study, participants reported how interesting they perceived the respective passage to be on a four-point Likert-like scale from 1 (*not interesting at all*) to 4 (*very interesting*). They were then given 4 min to either re-study the prose passage, answer the test questions with the aid of the materials (open-book test), or answer the questions by retrieving the initially studied information from memory (closed-book test). Ninety seconds before the end of the time limit, participants in the closed-book test condition received an answer sheet containing the solutions to the questions and were instructed to briefly self-check their answers without changing them (by circling correct answers and crossing out wrong answers). This served as brief feedback concerning the correctness of their given answers. Subsequent to the re-study condition, participants reported how often they were able to re-read the respective passage ($M = 1.63$, $SD = 0.67$; $N = 59$ due to missing data). After learning each passage, participants indicated the percentage of questions they expected to answer correctly in the final test (judgments of learning). After completing all three learning conditions, participants self-reported their prior knowledge concerning each of the three prose passages (from 0 to 100; see Appendix B for further analyses with participants' prior knowledge and for discussions of these).

The second session took place 2 weeks later. The final test consisted of three parts. Each part included the same seven short-answer test questions applied in the first session (in total: 21 final test questions). Participants worked for 6 min on each part of the final test. Their final test performance was operationalized as the proportion of correct answers per learning condition.

At the end of Session 2, participants answered demographic questions and control questions (e.g., if they had heard of the testing effect before and if they had studied the materials in the interim—92% of participants had not worked on the materials between Sessions 1 and 2). Finally, participants were debriefed and received their compensation.

3 | RESULTS

3.1 | Session 1: Practice test performance and judgments of learning

We conducted a paired-sample *t*-test to compare practice test performance—the proportion of correct answers given in the respective practice tests—when using open-book or closed-book tests in Session 1: Participants answered significantly more practice test questions correctly when working with an open-book test compared to working with a closed-book test, $t(61) = 7.41$, $p < .001$, $d_z = 0.94$ (see Table 1 for the respective descriptive statistics).

We further focused on participants' judgments of learning ($N = 61$ due to missing data): In general, participants expected to answer on average 33% of the final test questions correctly ($SD = 14.40$, range = 6.67–73.33). An analysis of variance (ANOVA) with repeated measures found no significant main effect of the learning condition on judgments of learning, $F(2, 120) = 1.61$, $p = .205$, $\eta_p = 0.03$. There were no significant differences between re-studying, open-book tests, and closed-book tests (see Table 1). Hence, participants did neither expect tests to be less nor more effective than re-studying.

TABLE 1 Descriptive statistics of practice test performances, final test performances, and judgments of learning (JOLs)

Learning condition	Performance		JOLs
	Practice test	Final test	
Re-study		.23 (.17)	32.10 (17.48)
Open-book test	.81 (.19)	.43 (.26)	34.75 (19.63)
Closed-book test	.51 (.23)	.48 (.26)	30.16 (16.55)

Note: Standard deviations of the means are displayed in parentheses. Practice test performances and final test performances are depicted in proportions of correct answers. JOLs, judgments of learning (in percentages).

3.2 | Session 2: Final test performance

Participants answered 38% of all 21 final test questions correctly ($SD = .12$, range = .14–.60). To test our hypotheses concerning final test performance, we conducted a repeated measures ANOVA: There was a significant main effect of the learning condition, $F(2, 122) = 17.88$, $p < .001$, $\eta_p = 0.23$. Pairwise comparisons with Bonferroni corrections indicated that using open-book tests and using closed-book tests yielded higher proportions of correct answers in the final test compared to re-studying (see Table 1 for the respective descriptive statistics; both $ps < .001$; $d_z = 0.66$, $d_z = 0.74$, respectively). We additionally conducted Bayesian analyses with SPSS 27 (using the default settings provided by SPSS): The two Bayesian *t*-tests respectively indicated extreme evidence favoring H1 (assuming that both open-book and closed-book tests lead to higher final test performances than re-studying) over H0 (assuming no difference between the respective learning conditions; both $BF_{01}s < 1/100$). A further pairwise comparison with Bonferroni corrections showed that using closed-book tests was similarly effective to using open-book tests ($p > .999$; $d_z = -0.12$). Here, the Bayesian *t*-test indicated moderate evidence for H0 assuming no difference between final test performances after open-book versus closed-book tests ($BF_{01} = 6.44$). Taken together, these findings supported our first hypothesis: both learning conditions including tests increased later final test performance compared to re-study. In contrast, our second, crucial, hypothesis was not supported: closed-book tests were not more beneficial than open-book tests—not even after a delay of 2 weeks.

Two additional paired-sample *t*-tests showed different degrees of forgetting in the two test conditions: While the high proportion of correct answers in the open-book tests in Session 1 could not be

maintained in the final test in Session 2, $t(61) = 12.35$, $p < .001$, $d_z = 1.57$, the proportion of correct answers initially given in the closed-book tests did not significantly differ from the proportion of correct answers achieved in the final test, $t(61) = 1.24$, $p = .221$, $d_z = 0.16$ (see Table 1).

4 | DISCUSSION

The present work was conducted to test the assumption that Agarwal et al.'s (2008) findings of similar learning benefits for open-book and closed-book tests was due to the rather short retention interval and that closed-book tests would be more beneficial than open-book tests with a longer final test delay. We focused on three learning conditions (re-study, open-book tests, and closed-book tests with feedback) based on those previously used by Agarwal and colleagues and increased the retention interval from 1 to 2 weeks.

The results of our work closely mirrored the findings of Agarwal et al. (2008) despite the longer delay: Both open-book and closed-book tests resulted in more correct answers in a final test than re-studying (even though the time limits given for initial study of the three prose passages as well as for answering the practice test questions were rather short). Thus, it seems relevant to further inform learners and educators about advantages of even short tests that could be easily applied at the end of school or university classes.

Contrary to our hypothesis but in line with the findings of Agarwal and colleagues, there was no difference between participants' final test performance following open-book or closed-book tests. Hence, both learning conditions including tests were similarly beneficial—even after a delay of 2 weeks. This supports the tentative conclusion based on Agarwal et al.'s (2008) findings that benefits of open-book and closed-book tests outweigh one another (see also, Agarwal & Roediger, 2011; Arnold et al., 2021). Initially, using open-book tests resulted in more correctly answered practice test questions in Session 1 compared to closed-book tests—which is unsurprising given that participants were able to consult the respective prose passages while answering open-book tests, whereas they had to retrieve the information from memory when using closed-book tests. However, this initial advantage of open-book tests did not persist over time: the proportion of correct answers in the final test after open-book tests was lower than the initial open-book test performance. In contrast, there was no decline between participants' initial closed-book test performance compared to the proportion of correct answers given in the final test, even with the longer delay of 2 weeks. This supports the conclusion from previous studies that retrieving information from memory leads to robust benefits.

Moreover, participants' judgments of learning did not differ between the three learning conditions and showed that participants overestimated the effectiveness of re-studying while underestimating the effectiveness of the practice tests. Interestingly, both open-book tests as well as closed-book tests were similarly underestimated even though open-book tests offered participants the opportunity to consult and re-study the materials when answering the practice test questions.

All in all, our findings support previous work concerning the benefits of applying tests as difficult learning tasks (e.g., Adesope et al., 2017; Rowland, 2014; Yang et al., 2021) and mirror the findings obtained by Agarwal et al. (2008). Hence, the results of Agarwal and colleagues seem to be robust and generalizable because we could replicate their findings even when using a slightly changed procedure, new learning materials, a different sample, and—most important—a longer delay between learning and the final test.

Although we were able to replicate the experiments of Agarwal et al. (2008) using their original paradigm with a longer final test delay, there are limitations of our work that we care to briefly discuss: For instance, our experiment was conducted in a laboratory using a rather small convenience sample, which might have been too small to validly test the critical comparison between participants' final test performances after using closed-book versus open-book tests. That is why we additionally conducted a Bayesian analysis—which, however, only resulted in moderate evidence for the null hypothesis. Given this still rather unsubstantial evidence and the generally small sample, further replications and future work comparing the effectiveness of open-book and closed-book tests with higher power (based on a-priori power analyses) are needed. In addition to such more highly powered studies, field experiments in actual schools or universities with more diverse samples are desirable (cf. Agarwal et al., 2021).

Apart from these limitations, we also want to briefly discuss ensuing ideas for future experiments. For instance, the test questions used in our work and in the original work of Agarwal et al. (2008) assessed mostly factual information given in the prose passages—thus, future experiments could explore if the here described findings are also applicable for transfer questions, inferences, or other questions particularly targeting higher-level thinking. Although previous work indicated that both closed-book and open-book tests are beneficial for a wide range of complex learning materials and for different types of test question formats (for overviews see e.g., Adesope et al., 2017; Yang et al., 2021), future work applying varying authentic and difficult materials as well as more elaborate and stimulating test questions would be valuable. Accordingly, especially in STEM courses like mathematics or physics, it is rather common to give learners, for instance, the possibility to consult formulae they can use to solve problems—which can be understood as an example of the utilization of open-book materials in applied settings. However, most until now conducted research only applied text-based learning materials (like textbook chapters, written notes, or prose passages) in open-book formats. It would thus be interesting to replicate our experiment and the experiments of Agarwal et al. (2008) using different types of materials in open-book formats (e.g., sheets including formulae, dictionaries, or reference works including definitions) to increase the generalizability of future research to authentic university or school settings. Most importantly, future work could focus more closely on learners' behavior while working on tests, because we cannot know for sure that participants answering open-book test questions did not engage in retrieval practice—or to what extend they retrieved the information. Although it seems obvious that they consulted the prose passages while answering the test questions (indicated by the difference

of correct answers between open-book and closed-book tests), it could be possible that participants first retrieved some (or all) information and then mostly used the prose passages as feedback to correct their answers. It would thus be advantageous to gather process data on how participants worked on the open-book and the closed-book tests or to directly manipulate to which degree (or when) learners have access to the materials while answering questions in an open-book test in future experiments. This would help to achieve a deeper understanding of the mechanisms underlying the beneficial effects of both open-book and closed-book tests. For instance, this applies to the important distinction of benefits due to difficult and effortful retrieval practice (as triggered by closed-book tests) and benefits due to more correctly answered test questions (as triggered by open-book tests). Finally, it seems important for future work to additionally contemplate whether even longer delays would change the here presented results concerning the similar benefits of open-book and closed-book tests. Although the implemented final test interval of 2 weeks represents a longer delay than the ones typically applied in laboratory experiments, and even though 2 weeks are a final test delay for which closed-book tests were demonstrated to be more effective than note-taking as a more elaborate control condition (Rummer et al., 2017), 2 weeks can still be seen as a rather short delay—especially in applied learning settings. To be able to transfer and generalize results and conclusions to schools or universities and to actual long-term learning effects, future experiments should thus focus on final test delays of multiple weeks or months in applied settings. Concluding, the present research was able to replicate the findings of Agarwal et al. (2008) and to show that their results remain robust even after a longer final test delay.

ACKNOWLEDGMENTS

We want to thank Marten Heuermann for help with data collection and data analysis as part of his bachelor's thesis. In addition, we thank Pooja Agarwal and an anonymous reviewer for valuable comments on an earlier version of this paper.

CONFLICT OF INTEREST

The authors have no relevant financial or non-financial interests to disclose.

ETHICS STATEMENT

The study was conducted in full accordance with the Ethical Guidelines of the German Association of Psychologists (DGPs) and the American Psychological Association (APA). By the time the data were acquired it was not customary at the respective University, or at most other German universities, to seek ethics approval for simple studies on learning or memory. Therefore, ethical approval was not required for this study in accordance with the national and institutional guidelines. Nonetheless, the study exclusively makes use of anonymous questionnaires. No identifying information was obtained from participants. Before starting, each participant had to provide their approval through reading and agreeing to a written informed consent. They were thereby explicitly informed that all data are treated confidentially and that they may withdraw from the study at any time without giving explanation.

DATA AVAILABILITY STATEMENT

The data of this study will be made openly available in OSF at https://osf.io/9vye4/?view_only=ed0e8018ae354c499c385641d8cbccb3.

ORCID

Kristin Wenzel https://orcid.org/0000-0002-0366-5222

REFERENCES

Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the use of tests: A meta-analysis of practice testing. *Review of Educational Research*, *87*(3), 659–701. https://doi.org/10.3102/0034654316689306

Agarwal, P. K., Karpicke, J. D., Kang, S. H. K., Roediger, H. L., III, & McDermott, K. B. (2008). Examining the testing effect with open- and closed-book tests. *Applied Cognitive Psychology*, *22*(7), 861–876. https://doi.org/10.1002/acp.1391

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*, *1-45*, 1409–1453. https://doi.org/10.1007/s10648-021-09595-9

Agarwal, P. K., & Roediger, H. L., III. (2011). Expectancy of an open-book test decreases performance on a delayed closed-book test. *Memory*, *19*(8), 836–852. https://doi.org/10.1080/09658211.2011.613840

Arnold, K. M., Eliseev, E. D., Stone, A. R., McDaniel, M. A., & Marsh, E. J. (2021). Two routes to the same place: Learning from quick closed-book essays versus open-book essays. *Journal of Cognitive Psychology*, *33*(3), 1–18. https://doi.org/10.1080/20445911.2021.1903011

Bamberg, E., Mohr, G., & Busch, C. (2012). *Arbeitspsychologie [occupational psychology]* (pp. 251–255). Hogrefe.

Batsell, W. R., Jr., Perry, J. L., Hanley, E., & Hostetter, A. B. (2017). Ecological validity of the testing effect: The use of daily quizzes in introductory psychology. *Teaching of Psychology*, *44*(1), 18–23. https://doi.org/10.1177/0098628316677492

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

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

Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. *From Learning Processes to Cognitive Processes: Essays in Honor of William K. Estes*, *2*, 35–67.

Buchin, Z. L. (2021). *Retrieval-based learning and element interactivity: The role of prior knowledge* (Doctoral dissertation, The University of North Carolina at Chapel Hill).

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

Dunlosky, J., Rawson, K. A., 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

Durning, S. J., Dong, T., Ratcliffe, T., Schuwirth, L., Artino, A. R., Boulet, J. R., & Eva, K. (2016). Comparing open-book and closed-book examinations: A systematic review. *Academic Medicine*, *91*(4), 583–599. https://doi.org/10.1097/ACM.0000000000000977

Ebersbach, M. (2020). Access to the learning material enhances learning by means of generating questions: Comparing open-and closed-book conditions. *Trends in Neuroscience and Education*, *19*, 100130. https://doi.org/10.1016/j.tine.2020.100130

Feller, M. (1994). Open-book testing and education for the future. *Studies in Educational Evaluation*, *20*(2), 235–238. https://doi.org/10.1016/0191-491X(94)90010-8

Hawkins, B. H. (2000). Buddhismus [Buddhism] (pp. 25–28). Herder.

Jacobs, L. C., & Chase, C. I. (1992). *Developing and using tests effectively: A guide for faculty*. Jossey-Bass Publishers.

Jensen, P. A., & Moore, R. (2009). Students' perceptions of their grades throughout an introductory biology course: Effect of open-book testing. *Journal of College Science Teaching*, *38*(3), 58–61.

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

Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborate studying with concept mapping. *Science*, *331*(6018), 772–775. https://doi.org/10.1126/science.1199327

Klaus, J. (1990). Alltag im "klassischen" Weimar [Everyday life in "classical" Weimar] (pp. 26–27). Nationale Forschungs- und Gedenkstätten der klassischen deutschen Literatur in Weimar.

McNamara, D. S., Kintsch, E., Songer, N. B., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction*, *14*(1), 1–43. https://doi.org/10.1207/s1532690xci1401_1

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

Richland, L. E., Bjork, R. A., Finley, J. R., & Linn, M. C. (2005). Linking cognitive science to education: Generation and interleaving effects. In *Proceedings of the twenty-seventh annual conference of the cognitive science society* (pp. 1850–1855). Erlbaum.

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

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

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

Rummer, R., & Schweppe, J. (2022). Komplexität und der Testungseffekt: Die mögliche Bedeutung der Verständnissicherung für den Nutzen von Abrufübung bei komplexem Lernmaterial [Complexity and the testing effect: The possible importance of securing comprehension for the benefit of retreival practice with complex learning material]. *Unterrichtswissenschaft*, *50*, 37–52. https://doi.org/10.1007/s42010-021-00137-4

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

Rummer, R., Schweppe, J., & Schwede, A. (2019). Open-book versus closed-book tests in university classes: A field experiment. *Frontiers in Educational Psychology*, *10*, 463. https://doi.org/10.3389/fpsyg.2019.00463

Waldeyer, J., Heitmann, S., Moning, J., & Roelle, J. (2020). Can generative learning tasks be optimized by incorporation of retrieval practice? *Journal of Applied Research in Memory and Cognition*, *9*(3), 355–369. https://doi.org/10.1016/j.jarmac.2020.05.001

Weissgerber, S. C., & Rummer, R. (2022). More accurate than assumed: Learners' metacognitive beliefs about the effectiveness of retrieval practice. Manuscript submitted for publication.

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

How to cite this article: Wenzel, K., Schweppe, J., & Rummer, R. (2022). Are open-book tests still as effective as closed-book tests even after a delay of 2 weeks? *Applied Cognitive Psychology*, *36*(3), 699–707. https://doi.org/10.1002/acp.3943

APPENDIX A

Further information concerning the applied materials

Example question applied in the practice test and in the final test.

Regarding the first prose passage, one example test question reads: What is allowed for Buddhist monks in Japan that is not allowed in other countries?

Correct answer: *Marriage*.

Descriptive statistics of the three prose passages depicted separately (Table A1).

APPENDIX B

Further analyses and discussion of the effects of prior knowledge

To test for influences of participants' prior knowledge on the effectiveness of the applied learning conditions, we first calculated a mean score of prior knowledge across all three-prose passages based on the three self-reports. Descriptively, participants reported rather low-prior knowledge across all materials ($M = 11.17$, $SD = 10.04$, range 0.00–43.33; $N = 60$ due to missing data). A conducted correlational analysis found a significant relation between participants' self-reported prior

TABLE A1 Descriptive statistics depicting participants' ratings of interestingness, judgments of learning, and prior knowledge concerning the three prose passages as well as their practice and final test performances dependent of the learning conditions

	Prose passage 1	Prose passage 2	Prose passage 3
Interestingness	2.82 (0.65)	2.28 (0.76)	2.76 (0.88)
Range	1.00–4.00	1.00–4.00	1.00–4.00
N	61	61	59
Judgments of learning	37.42 (19.98)	32.26 (17.60)	27.21 (18.00)
Range	10–70	0–90	0–70
N	62	62	61
Prior knowledge	7.33 (10.93)	10.17 (12.69)	16.00 (16.39)
Range	0–50	0–70	0–70
N	60	60	60
Practice test performance	.75 (.21)	.74 (.22)	.49 (.25)
Range	.29–1.00	.14–1.00	.00–1.00
N	41	42	41
Practice test performance—open-book	.90 (.13)	.87 (.13)	.65 (.21)
Range	.57–1.00	.64–1.00	.14–1.00
N	20	21	21
Practice test performance—closed-book	.61 (.18)	.60 (.22)	.32 (.17)
Range	.29–.93	.14–.95	.00–.71
N	21	21	20
Final test performance	.40 (.22)	.58 (.24)	.17 (.11)
Range	.00–.93	.00–.95	.00–.57
N	62	62	62
Final test performance—after open-book tests	.46 (.19)	.67 (.16)	.17 (.13)
Range	.14–.79	.43–.95	.00–.57
N	20	21	21
Final test performance—after closed-book tests	.53 (.18)	.70 (.17)	.20 (.11)
Range	.29–.93	.29–.90	.00–.36
N	21	21	20
Final test performance—after re-study	.20 (.14)	.35 (.19)	.13 (.09)
Range	.00–.43	.00–.64	.00–.29
N	21	20	21

Note: Standard deviations of the means are displayed in parentheses. Judgments of Learning and Prior Knowledge are depicted in percentages. The test performances are depicted in proportions of correct answers.

knowledge and their performance in the final test ($r = .26$, $p = .045$). Hence, we conducted an analysis of covariance with repeated measures to exploratively assess potential effects of participants' (*z*-standardized) prior knowledge on the effectiveness of the learning conditions: There was a significant main effect of the learning condition on final test performance, $F(2, 116) = 17.98$, $p < .001$, $\eta_p = 0.24$. Subsequent pairwise comparisons with Bonferroni corrections indicated that open-book tests and closed-book tests were more beneficial than re-studying (both *p*s < .001) but that the effectiveness of closed-book tests and open-book tests did not significantly differ from each other ($p > .999$). Moreover, there was a significant main effect of prior knowledge on final test performance, $F(1, 58) = 4.19$, $p = .045$, $\eta_p = 0.07$. The interaction of the learning condition and participants' prior knowledge was also significant, $F(2, 116) = 4.06$, $p = .020$, $\eta_p = 0.07$. Subsequent parameter estimates indicated that prior knowledge had neither significant effects on final test performance after re-study, $t(116) = 0.04$, $B = 0.001$, SE = 0.02, $p = .969$, nor after open-book tests, $t(116) = -0.42$, $B = -0.02$, SE = 0.04, $p = .675$, but had a significant positive effect on final test performance after closed-book tests, $t(116) = 3.30$, $B = 0.10$, SE = 0.03, $p = .002$.

Summarizing, prior knowledge did neither moderate the effects of re-studying nor of open-book tests but increased the benefits of closed-book tests. This fits the assumption that positive effects of tests arise because taking tests and retrieving information leads to connections of the retrieved information and information already stored in memory as well as the assumption that higher prior knowledge is required to be even able to solve difficult (retrieval) tasks and to benefit from them (e.g., Bjork & Bjork, 1992, 2011). In contrast, it is also possible that taking practice tests in an open-book fashion compensated for lower prior knowledge and thus allowed these participants to benefit from the test questions more than when they had to answer them from memory.

Hence, participants' prior knowledge proved to be important for the effectiveness of closed-book tests—even though we used an aggregated prior knowledge score and even though participants only briefly self-reported their prior knowledge after completing all learning conditions. Hence, future work could assess prior knowledge more objectively and more thoroughly (e.g., by implementing prior knowledge tests), which might be even more predictive. In line with this, future work should assess prior knowledge before (and not after) participants start to work on the learning conditions—because especially working on test questions (and getting feedback in the closed-book test condition) might have distorted participants' perceptions and ratings of their prior knowledge.