Psychon Bull Rev (2015) 22:135–140
DOI 10.3758/s13423-014-0646-x

BRIEF REPORT

Retrieval practice: the lack of transfer to deductive inferences

Randy Tran · Doug Rohrer · Harold Pashler

Published online: 17 May 2014

Abstract Retrieval practice has been shown to enhance later recall of information reviewed through testing, whereas final-test measures involving making inferences from the learned information have produced mixed results. In four experiments, we examined whether the benefits of retrieval practice could transfer to deductive inferences. Participants studied a set of related premises and then reviewed these premises either by rereading or by taking fill-in-the-blank tests. As was expected, the testing condition produced better final-test recall of the premises. However, performance on multiple-choice inference questions showed no enhancement from retrieval practice.

Keywords Testing effect · Transfer of learning · Retrieval practice · Deductive inference

Introduction

A great many studies have found that review through testing produces better final recall than does review through restudying. This advantage, often dubbed the *testing effect* or the *retrieval practice effect*, has been studied for many years (e.g., Carrier & Pashler, 1992; McDaniel, Anderson, Derbish, & Morrisette, 2007; Roediger & Karpicke, 2006; Tulving, 1967). However, the scope of testing benefits has not been fully charted.

The testing effect

Robust benefits of retrieval practice have been found in a variety of memory tasks, including free recall of word lists (e.g., Tulving, 1967), paired-associate learning (e.g., Carpenter, Pashler, & Vul, 2006), foreign language vocabulary learning (e.g., Carrier & Pashler, 1992), and learning content from prose passages (e.g., Roediger & Karpicke, 2006). Testing effects have also been found in the classroom. For example, Carpenter, Pashler, and Cepeda (2009) showed that 8th grade students learning factual information (e.g., *Who assassinated President Abraham Lincoln?*) performed better for items on a memory test 9 months later when the review took the form of testing rather than restudy (see also McDaniel, Anderson, et al., 2007; McDaniel, Roediger, & McDermott, 2007). Although the testing effect appears robust to changes in material and setting, the great majority of the testing effect studies have focused on explicit retrieval of the same information that was reviewed through testing.

Transfer of the testing effect

Does testing facilitate learning, as measured by the learner's ability to draw conclusions going beyond the information studied? In one of the few studies on this question, Rohrer, Taylor, and Sholar (2010) had children learn maps through study only or cued recall. In their first study, for example, participants in the retrieval practice condition were shown map regions, one at a time, and were asked to supply the name of the region. Cued recall produced greater scores on a final transfer test requiring the children to both freely recall each region's name and identify its location. In another study, McDaniel, Anderson, et al. (2007) found transfer of the testing

Electronic supplementary material The online version of this article (doi:10.3758/s13423-014-0646-x) contains supplementary material, which is available to authorized users.

R. Tran (✉) · H. Pashler
Department of Psychology, University of California, San Diego, 9500 Gilman Drive MC0109, La Jolla, CA 92092-0109, USA
e-mail: r4tran@ucsd.edu

D. Rohrer
Department of Psychology, University of South Florida, Tampa, FL, USA

effect when participants had to recall a previously unretrieved keyword on a final test using the same sentence that included a previously retrieved keyword. However, the tests used in these studies arguably required only limited transfer (Barnett & Ceci, 2002).

In a recent study explicitly designed to assess far transfer, Butler (2010) used a final test involving inference questions about the material learned. The final assessment required participants to answer inference questions by applying the knowledge learned to a topic within a similar domain (e.g., *Sometimes bats die while they are sleeping. What will happen if a bat dies while it is hanging upside down?*) or a different domain (e.g., *The U.S. Military is looking at bat wings for inspiration in developing a new type of aircraft. How would this new type of aircraft differ from traditional aircrafts like fighter jets?*).

Butler (2010) found better performance on inference test questions for items that had been tested. However, the final inference questions used in Butler's study were not clear-cut examples of either deductive inferences (i.e., drawing a logically necessary conclusion from a set of premises) or inductive inferences (i.e., generalizing from multiple examples). For example, Butler's participants read the following statement about bread: "*unleavened bread has symbolic importance in many religions and, thus, nowadays it is primarily consumed in the context of religious rites and ceremonies.*" and were then asked "*Roman Catholic Christians use bread when they celebrate the Eucharist, a rite derived from the narrative of the Last Supper. What type of bread is likely to be used in this religious ceremony?*" The cue for recall of unleavened bread stems from "*rite*" and "*religious ceremony.*" While it is clear why a reader might volunteer "unleavened" given that the passage contained no other relevant information, this conclusion would not seem to be either a valid deductive or even a valid inductive inference. Other inference questions seemed to us similarly ambiguous—invited by the passage but not logically warranted by it. Thus, it seemed conceivable that participants might have interpreted the test as a cued recall test, asking themselves, in effect, "what information was contained in the passage that might be relevant to this question?" If so, the occurrence of a testing effect might occur for the same reason as the testing effect in cued recall, whether or not testing facilitates inferences.

Present study

To shed further light on the effect of retrieval practice on making inferences, we created a set of learning materials and test questions that specifically required transfer to deductive inferences, by which we mean drawing a conclusion that depends logically on multiple premises, each of which was learned in isolation.

Here, we report four experiments asking whether the benefits of retrieval practice extend to inference questions. The learning material consisted of four *scenarios*, each composed of seven to nine facts or *premises*. For each scenario, participants completed a presentation, learning, and assessment phase. In the presentation phase, premises of a scenario were presented sequentially and only once. Then participants studied those premises by either rereading or retrieving missing keywords(s) of the premises. After each learning phase of a scenario, participants were assessed with a final test consisting of eight multiple-choice deductive inference questions. In some of the experiments, we also manipulated the retention interval by testing participants immediately or 48 h after the learning phase. This allowed us to examine the efficacy of the testing effect on deductive inferences with a longer delay.

Experiment 1

In Experiment 1, all participants were assessed immediately after the learning phase.

Method

Participants Sixty-eight undergraduates at the University of California, San Diego participated in this experiment for course credit. All were naïve as to the purpose of the experiment.

Materials Four unrelated scenarios were created with seven to nine premises each. Each premise within a scenario shared a common theme, but all were logically independent. Together, though, each set of premises had a number of logical implications, which were assessed by the multiple-choice transfer test (i.e., deductive inference questions). All the materials are listed in Supplementary Online Materials.

Design A two-level single-factor within-subjects design was used. During the learning phase, participants reviewed the premises by either rereading or retrieval practice. The two study conditions were counterbalanced across the four scenarios.

Procedure Participants were tested individually in sound-attenuated booths for the computerized study. The entire experiment, including consent and debrief, was completed within a single 1-h session.

Presentation phase Participants were instructed to read the sequentially presented premise. They were also told they would later be required to make inferences from the studied premises.

Learning phase After the initial presentation phase, participants reviewed the premises by one of two methods: reread or retrieval practice. Participants were told that they had 5 min to cycle through the premises at their own pace and that clicking quickly would not decrease the duration of the learning phase. For the duration of the learning phase, premises were presented in blocks and randomized within each block. In the *reread condition*, participants were instructed to reread each premise and click "continue" for the next premise to be presented. In the *retrieval practice condition*, participants were instructed that each premise would have missing keyword(s) and that they should covertly recall the missing word (i.e., retrieve the missing word either silently or aloud, without recording their response) before clicking "continue." Then the correct completed premise would appear. Finally, participants clicked "continue" to be presented with the next premise.

Assessment phase Immediately after the learning phase, participants were given eight multiple-choice questions that required them to make inferences using the learned premises. Each of the inference questions required information from at least two of the premises to be answered correctly. For example, participants studied the premises—(1) "*The local dealership has 7 Calientes on the lot.*" (2) "*At your local dealership, most of the Calientes have the 4-cylinder engine.*" (3) "*At your local dealership, every 4-cylinder Caliente is black.*"—and were then asked, "*What is the smallest possible number of black Calientes on the lot?*"

All three phases were repeated for the subsequent scenarios.

Results and discussion

Overall, performance (M = 0.78, SEM = 0.02) was well above chance (chance = 0.24), $t(67)$ = 30.55, $p < .001$, 95 % CI [0.75, 0.82]. However, performance was not significantly different between the reread condition (M = 0.80, SEM = 0.02) and the retrieval practice condition (M = 0.77, SEM = 0.02), $t(67)$ = 1.35, p = .18. The effect size was g^* = 0.16 favoring rereading, 95 % CI [−0.08, 0.40]. Experiment 1 demonstrated that there was no benefit of retrieval practice on the final inference assessment. This lack of an effect, however, could be due to the use of an immediate assessment, because testing effects are generally found after a delay between study and assessment (see Roediger & Karpicke, 2006).

Experiment 2

To examine whether the lack of transfer of the testing effect in Experiment 1 was due to the absence of a test delay, all participants in the second study were assessed after a 48-h delay. Otherwise, Experiment 2 was the same as Experiment 1.

Method

Participants Forty participants from the same population that was used in Experiment 1 participated in this experiment for course credit. All were naïve as to the purpose of the experiment.

Materials Materials were identical to those used in Experiment 1.

Design The design of Experiment 2 was identical to that of Experiment 1.

Procedure The procedure was identical to that of Experiment 1, with the following exception: Instead of completing the entire experiment in a single session, participants completed the learning phase during one session and returned 48 h later to complete the assessment.

Results and discussion

Again, we found no benefit of retrieval practice on inference questions. As with Experiment 1, participants in Experiment 2 performed above chance (M = 0.54, SEM = 0.03), $t(39)$ = 10.05, $p < .001$, 95 % CI [0.48, 0.61]. The assessment performances on inferences for rereading (M = 0.55, SEM = 0.04) and retrieval practice (M = 0.54, SEM = 0.03) were nearly identical with no significant difference, $t(39)$ = 0.36, p = .72, g^* = 0.06, 95 % CI [−0.25, 0.37]. After debriefing and interviewing participants, a potential reason for the lack of an effect could be that participants may not have carried out retrievals in the retrieval condition. In both Experiments 1 and 2, participants in the retrieval practice conditions were asked to recall the missing words covertly rather than overtly. Although Smith, Roediger, and Karpicke (2013) found both covert and overt retrieval produced equally large testing effects, their covert retrieval procedure gave participants a fixed duration of 40 s to covertly retrieve the learned information. Our learning phase, however, was self-paced. Participants could have simply clicked "continue" to obtain the correctly filled premise, without first trying to retrieve the missing keyword, essentially simulating the reread condition.

Experiment 3

The results of Experiments 1 and 2 showed no testing effect. However, we were uncertain whether participants in the retrieval practice condition completed the task by covertly

retrieving the missing words in each premise, as instructed. In the retrieval practice condition in the present experiment, participants were asked to type the missing keywords before seeing the correct response. Finally, unlike in Experiments 1 and 2, we manipulated the duration of the test delay (0 vs. 48 h).

Method

Participants One hundred seventy-three participants from the same population participated in this experiment for course credit. Five participants were excluded from the analysis for failing to comply with the instructions, leaving 84 participants in each of the two groups (0- or 48-h test delay). All were naïve as to the purpose of the experiment.

Materials Materials were identical to those in Experiments 1 and 2.

Design A 2 × 2 design was used. The studying condition (reread vs. retrieval practice) was manipulated within subjects and counterbalanced across the four scenarios. The retention interval was manipulated between-subjects (immediate or 48 h).

Procedure The procedure was identical to that of Experiments 1 and 2, with the following exception: In the retrieval practice condition, participants were required to type, rather than covertly recall, the keywords missing from the premises.

Results and discussion

The results were similar to the results of Experiments 1 and 2. There was a main effect of test delay, $F(1, 166) = 49.74$, $p < .001$, $\eta_p^2 = .23$, reflecting a drop from 72.02 % (*SEM* = 1.42 %) to 53.94 % (*SEM* = 1.52 %). However, there was no effect of study condition, $F(1, 166) = 1.49$, $p = .22$, $\eta_p^2 = .01$, and no interaction, $F(1, 166) = 0.84$, $p = .36$, $\eta_p^2 = .01$ (see Fig. 1). In brief, we once again found no testing effect on transfer.

Some informal interviewing of participants in Experiment 3, along with some reflection, led us to two possible explanations of our inability to find a testing effect on transfer tasks. (1) Retrieval practice on the current stimuli set might not have produced the typical testing effect found in recalling individual pieces of information. (2) In the retrieval practice condition, participants were asked to recall the same missing words each time they saw a particular premise. Therefore, a participant might have superficially attended only to the blanks—using the peripheral words as cues—without fully processing each premise (Hinze & Wiley, 2011).

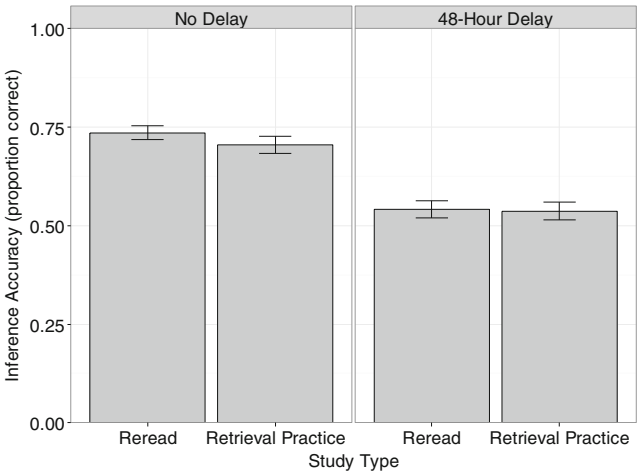

Fig. 1 Mean proportion correct on multiple-choice inference questions immediately after study or after a 48-h delay as a function of the study condition (reread vs. retrieval practice) in Experiment 3. Error bars represent standard errors of the means. Chance accuracy was 24 %

Experiment 4

In Experiment 4, we sought to determine whether or not the typical benefits of retrieval practice could be found using our specific set of stimuli. In addition, to increase the likelihood that participants in the retrieval practice condition attended to the entire premise before trying to recall the missing keywords, we created multiple versions of each, and the missing keywords varied across versions. Finally, given that the size of the testing effect typically increases with delays between study and final assessment (e.g., Roediger & Karpicke, 2006), we delayed the test until 48 h after study.

Method

Participants One hundred sixty-eight participants from the same population participated in this experiment for course credit. Four participants were excluded from the analysis for failing to comply with the instructions, leaving 80 participants for the fill-in-the-blank assessment and 84 participants for the inference assessment. All were naïve as to the purpose of the experiment.

Materials The same scenarios and premises were used in this experiment; however, multiple versions with different keywords missing were created and used for the retrieval practice.

Design Half of the participants received a fill-in-the-blank final test on the premises, and the other half received a multiple-choice inference test. For each group, study strategy (reread or retrieval) was manipulated within subjects. In the retrieval practice condition,

the keywords missing from each premise varied across blocks.

Procedure The procedure was identical to that of Experiment 3, with the following exception: All participants were assessed after 48 h and on one of two different final tests. The two final assessments were analyzed separately.

Results and discussion

Fill-in-the-blank performance Two coders blind to the conditions scored the fill-in-the-blank answers. Interrater reliability was high, with a Pearson's r of .99; therefore, we randomly chose one coder, and only that coder's scores were used for the subsequent analysis. Consistent with findings in the literature, we observed a significant difference between the retrieval practice condition ($M = 0.87$, $SEM = 0.01$) and reread condition ($M = 0.77$, $SEM = 0.02$) when participants were assessed on filling-in-the-blank, $t(79) = 5.34$, $p < .001$, $g^* = 0.59$, 95 % CI [0.35, 0.83] (see Fig. 2). This confirmed that our retrieval practice condition produced the typical testing effect with the stimuli used in all of the studies presented here.

Inference performance Again, we found no significant performance difference between the retrieval practice condition ($M = 0.50$, $SEM = 0.02$) and reread condition ($M = 0.53$, $SEM = 0.02$) on the final inference assessment, $t(83) = 1.75$, $p = .08$, $g^* = 0.19$, 95 % CI [−0.03, 0.40] (see Fig. 3). Our results replicated Experiment 2 and the 48-h delay condition in Experiment 3 even when the missing keywords for each premise varied. Given that we found the typical testing effect when participants were assessed on recall for the missing keywords, we concluded that the benefits of retrieval practice do not extend to making deductive inferences.

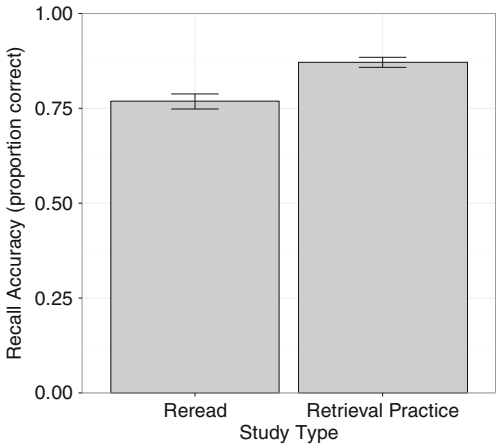

Fig. 2 Mean proportion correct on recall on the version with the most missing keywords for each premise after a 48-h delay in Experiment 4. Error bars represent standard errors of the means

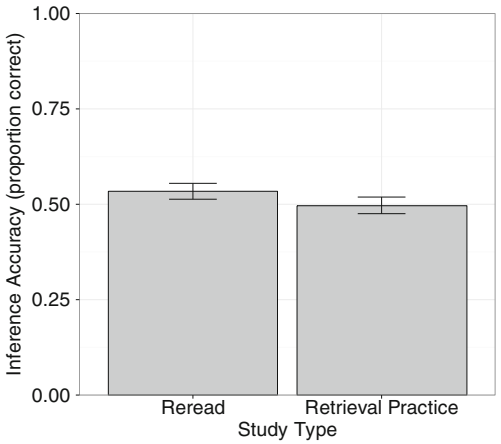

Fig. 3 Mean proportion correct on multiple-choice inference questions after a 48-h delay as a function of the study condition (reread vs. retrieval practice) in Experiment 4. Error bars represent standard errors of the means. Chance accuracy was 24 %

General discussion

In the present study, we asked whether the benefits of retrieval practice review (answering fill-in-the-blank questions) transfer to solving deductive inference questions based on the content reviewed.

In four experiments, participants learned various premises relating to four fictional scenarios and were later asked to make deductive inferences that depended upon these premises. Participants reviewed half of the scenarios by rereading, and the other half by covertly recalling or typing the missing keywords from each premise. Despite our various efforts to modify the methods in ways that might elicit retrieval practice benefits, we found no gains in our transfer tests (and some trends in the other direction). However, Experiment 4 confirmed that testing enhanced recall for the premises we were using, as expected (see also Supplementary Online Methods for results of a control experiment measuring inference performance when premises were provided during inference test).

Why was there no transfer?

The format of the retrieval practice condition required participants to retrieve missing keywords for each premise. As was noted above, it seemed possible that in the retrieval practice condition, participants might have superficially attended to the fixed blanks and relied on the location of omitted portions of the premise as retrieval cues (Hinze & Wiley, 2011). In Experiment 4, the keywords missing from each premise varied from trial to trial to prevent this sort of superficial processing, but the results were unchanged.

The lack of a testing benefit on transfer superficially contrasts with the results of Butler (2010). Why the different outcome? The present study presented each premise sequentially, possibly encouraging participants to complete the task

of retrieving the missing keywords without much regard to how the premises were related during the retrieval practice condition. Unlike the retrieval practice procedure used in the present experiments (i.e., fill-in-the-blank task), Butler's retrieval practice procedure involved answering short answer questions. Hinze and Wiley (2011) demonstrated that performance on a final test consisting of novel multiple-choice questions was higher when review involved answering more open-ended questions (e.g., short answer), as compared with a fill-in-the-blank task, possibly requiring more integration. It is also possible, as was noted earlier, that Butler's materials evoked task-specific strategies due to the fact the inferences lacked logical necessity.

Another question that may occur to the reader is: Exactly how is it that retrieval practice enhanced recall of the premises but did not enhance inference making? If learners were able to recall more premises, how could this fail to improve inference performance? For the retrieval practice condition, participants must actively recall multiple premises and check whether the premises recalled are relevant to the presented inference question (i.e., item-specific processing). By contrast, for participants in the rereading condition, the lack of such demands may have allowed them time and resources to attend "online" to the relationships between premises (i.e., relational processing). This cognitive work in the learning phase may have paid special dividends in the inference task, but not in the explicit memory task.

To better understand the difference between our findings and Butler's (2010), it might be useful to perform an experiment where the premises of the four scenarios are one at a time, with the scenarios interleaved. Presenting the mixed premises would likely make it difficult for participants to examine the relationships between premises during the learning phase. Therefore, we would expect that participants would not be able to draw inferences in the rereading condition leading to a decrease on inference performance, whereas inference performance in the retrieval practice condition should be unaffected for reasons discussed previously.

It is also worth noting that the multiple-choice inference test might somehow have been insensitive to performance differences between conditions. Although our test questions were designed not to be answerable on the basis of mere familiarity, it is always possible that our materials did not achieve this goal for some reason.

Limitations and directions for future research

On the basis of the present results, it appears that although retrieval practice robustly facilitates explicit learning of factual material, this does not always improve flexible use of information. Of course, the form of retrieval used here (i.e., fill-in-the-blank) was only one of the many potential forms of retrieval that could be examined as methods of review (others would involve short-answer, multiple-choice, or free recall questions). Testing is undoubtedly a useful technique for promoting information acquisition, but we need to know more about when it does and when it does not facilitate transfer of learning.

Acknowledgments This work was supported by a collaborative activity award to H. Pashler from the J. S. McDonnell Foundation, a MURI award from the Office of Naval Research (25684A), and an NSF Grant (SBE-0542013, G.W. Cottrell, PI). We thank Noriko Coburn for many thoughtful comments and Dorothy Uong, Monica Kullar, Matt Su, and Alison Bennett for collecting and/or scoring the data. We would also like to thank Andy Butler and Roddy Roediger for providing many helpful comments on a previous version of the manuscript.

References

Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn?: A taxonomy for far transfer. *Psychological Bulletin, 128*(4), 612–637. doi:10.1037/0033-2909.128.4.612

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. doi:10.1037/a0019902

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

Carpenter, S. K., Pashler, H., & Vul, E. (2006). What types of learning are enhanced by a cued recall test? *Psychonomic Bulletin & Review*, 13(5), 826–830. doi:10.3758/BF03194004

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition, 20*(6), 633–642. doi:10.3758/BF03202713

Hinze, S. R., & Wiley, J. (2011). Testing the limits of testing effects using completion tests. *Memory, 19*(3), 290–304. doi:10.1080/09658211.2011.560121

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

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

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

Rohrer, D., Taylor, K., & Sholar, B. (2010). Tests enhance the transfer of learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(1), 233–239. doi:10.1037/a0017678

Smith, M. A., Roediger, H. L., & Karpicke, J. D. (2013). Covert retrieval practice benefits retention as much as overt retrieval practice. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*(6), 1712–1725. doi:10.1037/a0033569

Tulving, E. (1967). The effects of presentation and recall of material in free-recall learning. *Journal of Verbal Learning and Verbal Behavior, 6*(2), 175–184. doi:10.1016/S0022-5371(67)80092-6