2010

Long-term Effects of Testing on the Recall of Nontested Materials

Jason C.K. Chan
Iowa State University, ckchan@iastate.edu

Long-term Effects of Testing on the Recall of Nontested Materials

Abstract

Testing, or memory retrieval, is a powerful way to enhance long-term retention of studied material. Recent studies have shown that testing can also benefit later retention of related but nontested material (a finding known as retrieval-induced facilitation, Chan, McDermott, & Roediger, 2006), but the long-term consequences of this benefit is unknown. In the current experiment three retention intervals—20 minutes, 24 hours, 7 days—were used to assess the effects of testing on subsequent recall of the nontested material. The results indicate that the magnitude of retrieval-induced facilitation, like that of the testing effect (i.e., the memorial benefit of testing on the tested material), increases with delay at the beginning (i.e., between 20 minutes and 24 hours) but asymptotes afterward (i.e., between 24 hours and 7 days). Theoretical and applied implications of this finding are discussed.

Disciplines

Cognitive Psychology | Educational Psychology

Comments

This is an Accepted Manuscript of an article published by Taylor & Francis in Memory on 2010, available online: http://www.tandf.com/10.1080/09658210903405737

Running Head: LONG-TERM EFFECTS OF TESTING ON NONTESTED MATERIALS

Long-term Effects of Testing on the Recall of Nontested Materials

Jason C.K. Chan

Iowa State University of Science and Technology

Correspondence:

Jason C.K. Chan
Department of Psychology
W112 Lagomarcino Hall
Ames, IA 50011-3180
USA
ckchan@iastate.edu

Abstract

Testing, or memory retrieval, is a powerful way to enhance long-term retention of studied

material. Recent studies have shown that testing can also benefit later retention of related, but

nontested material (a finding known as retrieval-induced facilitation, Chan, McDermott, &

Roediger, 2006), but the long-term consequences of this benefit is unknown. In the current

experiment, three retention intervals – 20 min, 24 hr, 7 days – were used to assess the effects of

testing on subsequent recall of the nontested material. The results indicate that the magnitude of

retrieval-induced facilitation, like that of the testing effect (i.e., the memorial benefit of testing

on the tested material), increases with delay at the beginning (i.e., between 20 min and 24 hr) but

asymptotes afterward (i.e., between 24 hr and 7 days). Theoretical and applied implications of

this finding are discussed.

The Long-term Effects of Testing on Recall of the Nontested Materials

Retrieval is a potent memory enhancer. Over the past century, research has repeatedly demonstrated the robust benefits of testing on retention (Abbott, 1909; Chan & McDermott, 2007; Karpicke & Roediger, 2008; Spitzer, 1939). Recently, researchers have suggested that frequent testing should be implemented in the classroom to promote learning (Bangert-Drowns, Kulik, & Kulik, 1991; Chan et al., 2006; Leeming, 2002; McDaniel, Anderson, Derbish, & Morrisette, 2007; McDaniel, Roediger, & McDermott, 2007; Roediger & Karpicke, 2006b). Further establishing testing as a beneficial learning tool is the finding that testing not only promotes retention of the tested material but also that of the nontested material (so long as two are related to each other, Chan et al., 2006). For example, if a student has learned two related concepts such as "the wavelength emitted by something travelling away from us is shifted to a *lower* frequency" and "the visible wavelengths emitted by objects that are moving away from us are said to be redshifted," recalling the first fact during an initial test can facilitate subsequent recall of the second fact, even though the second fact was not tested initially. This phenomenon is termed retrieval-induced facilitation (henceforth RIFA).

Although RIFA is a relatively new finding, the evidence that testing can enhance later recall of nontested materials is not without precedence (Callender & McDaniel, 2007; Carpenter, Pashler, & Vul, 2007; Cranney, Ahn, McKinnon, Morris, & Watts, 2009; Hamaker, 1986; Pilotti, Chodorow, & Petrov, 2009). Notably, testing can also impair later recall of the nontested materials – a phenomenon termed retrieval-induced forgetting (henceforth RIFO). In a typical RIFO experiment, participants first learn a series of category-exemplar pairs (e.g., furniture – chair, furniture – table, drinks – whisky, drinks – scotch). They then are tested on half of the exemplars from half of the categories. After a short delay (usually about 20 min), participants

are tested on all the pairs. The general finding is that initial testing on some of the exemplars can impair later recall of their related exemplars.

The RIFO literature has been thoroughly reviewed by Anderson (2003), and the relation between RIFO and RIFA is covered elsewhere (Chan, 2009). Briefly, two factors – integration during encoding and delay between the initial and the final test – appear to be important determinants of whether initial testing facilitates or impairs later recall of nontested information. RIFO is typically eliminated (1) if participants integrate the studied material into a coherent, inter-connected whole (e.g., Anderson, Green, & McCulloch, 2000; Bauml & Hartinger, 2002; Migueles & Garcia-Bajos, 2007) and (2) if there is a substantial retention interval (i.e., 24 hr or more) separating the initial and final tests (e.g., Bjork, Bjork, & MacLeod, 2006; Chan et al., 2006). When integration is combined with a long delay, testing can facilitate recall of the nontested materials (Chan, 2009). Note that the current paper focuses not on the distinction between RIFO and RIFA, but rather on the interactive effects of delay and testing on the retention of initially nontested information.

According to Chan et al. (2006), RIFA occurs because people actively search for information related to the target memory when they attempt to answer a question (cf. Burgess & Shallice, 1996). This active search may occur because people have trouble coming up with the target memory initially or because they have spare time to retrieve additional information after they have already recalled the target. There are two importance assumptions to this explanation. First, RIFA and the testing effect may share the same underlying mechanism – an active and conscious retrieval process that enhances retention of the retrieved memories, regardless of whether such memories were retrieved because they were directly tested or were retrieved in a collateral fashion (as in RIFA, cf., Kornell, Hays, & Bjork, 2009). Second, RIFA can be

attributed mostly to a controlled, instead of an automatic (e.g., spreading activation), process. As a result, they should be strengthened in the long run (unlike semantic priming), much like the directly tested facts in the testing effect.

In the current experiment, I sought to examine the effects of testing on the forgetting of the tested and nontested information. There are three important reasons to examine the effects of initial testing on subsequent recall at various retention intervals. First, from an applied perspective, it is important to know whether RIFA occurs in the long run, because so far RIFA in episodic memory has only been shown at delays lasting up to 24 hr (Chan, 2009; Chan et al., 2006). Indeed, other related memory phenomena have been shown to vary with retention intervals. For example, the effects of retrieval inhibition and expanded retrieval practice have both been suggested to vary with delay (Bjork et al., 2006; Karpicke & Roediger, 2007). Therefore, if one wants to argue that testing enhances learning of both the tested and nontested materials from an educational perspective (e.g., Carpenter et al., 2007; Chan et al., 2006), then it is important to show that RIFA extends beyond a mere one-day delay.

Second, from an empirical perspective, the influence of long-term delay on the recall of the nontested materials is not clear. While some studies have shown that testing can enhance the delayed recall of nontested-related material relative to no initial testing (e.g., Chan et al., 2006; Pilotti et al., 2009), others have reported the contrary (e.g., Conroy & Salmon, 2006; Garcia-Bajos, Migueles, & Anderson, 2009; Tandoh & Naka, 2007). For example, Tandoh and Naka showed that testing can impair delayed recall of the nontested materials, but they did not used integrative, textbook like materials. On the other hand, while Conroy and Salmon (2006) used coherent materials, they did not manipulate retention interval (i.e., delay was held at 1 day). Garcia-Bajos and colleagues (2009) used a video event and manipulated the retention interval

between the initial and final test (immediate and 1-week delay). These authors found that highly integrated materials showed no RIFO, but the poorly integrated materials showed RIFO even after one week. Therefore, the current experiment sought to further investigate the relation between delay and RIFA/RIFO. One important feature of the present study is that three retention intervals were examined, such that the influence of testing on the forgetting of the notested materials can be observed in both the short- and long-term (see Slamecka & Katsaiti, 1988, for a discussion of the importance in examining three or more retention intervals).

The third reason to examine the long-term effects of testing is that previous research has demonstrated conflicting findings about whether testing slows forgetting of the tested materials. For example, some researchers argue that testing retards forgetting of the tested materials (e.g., Carpenter, Pashler, Wixted, & Vul, 2008; Roediger & Karpicke, 2006a) whereas others claim the opposite (Slamecka & Katsaiti, 1988). Although considerable research has examined the effects of testing on the forgetting of the tested materials, little is known about the nontested materials.

<div align="center">Method</div>

Design

The experiment contained three phases. In the first phase, participants read two articles (one on the Shaolin Buddhist temple and one on the Big Bang theory) and took an immediate recall test on one of the articles twice (i.e., they completed the same test back-to-back). During the second phase, participants completed two filler tasks, including the computerized Operation Span task (OSPAN) (Conway et al., 2005) and a mental arithmetic task for a total of 20 min. During the third phase, participants completed the final test for both the tested and nontested articles. The initial test contained 12 questions, but the final test contained 24 questions (for each topic). This procedure thus created three question sets: a tested set, a nontested-related set

(i.e., nontested questions from the tested article), and a control set (i.e., questions from the nontested article). The experiment used a mixed 3 (delay: 20 min, 24 hr, 7 days) X 3 (question type: tested, nontested-related, control) design. Delay was manipulated between-subjects and question type was manipulated within-subjects.

Participants

A total of 84 students from Washington University and Iowa State University participated in this experiment for course credits or for $20. There were 28 participants in each delay condition. They were tested individually or in groups of up to 5 people. The experimenter posted research opportunities on a webpage and participants signed up for their preferred study. All participants were given the option of receiving course credit or payment. Despite the non-random nature of participants assignment into delay conditions, no group differences were found based on the initial test results as will be revealed later in the results section.

Materials

All materials were the same as those used in Experiments 2 and 3 of Chan et al. (2006). Both the Shaolin article and the Big Bang article were about 1900 words long. Two related sets of 12 questions were created for each article. For example, one question from the Big Bang article was, "Due to Doppler shifting, the wavelength emitted by something traveling away from us is shifted to a _____ frequency." The answer to this question was "lower." Its related question in the other set was, "When visible wavelengths are emitted by objects moving away from us, their wavelengths are said to be _____ (Name the phenomenon)." The answer to this question was "redshifted." The questions were analyzed for relatedness with the sentence comparison feature in latent semantic analysis (Foltz, Kintsch, & Landauer, 1998), which indicated that the Big Bang questions and the Shaolin

questions had a relatedness rating (a correlation) of .39 and .26, respectively. These relatedness

ratings are higher than if the questions within each article are paired randomly (.18 and .10 for

Big Bang and Shaolin, respectively), both ts > 2.74, ds > 1.11. For the complete set of questions

and their answers, see Appendix C of Chan (2009).

Procedure

During Phase 1 of the experiment, participants were given 16 minutes to study each

article and were told that they might take a test on none, one, or both of the articles immediately

after the encoding phase (in reality, they took the initial test for only one article). No explicit

instructions to integrate the study materials were given. Whether participants read the tested or

nontested article first was counterbalanced across participants. During the initial test, questions

appeared on the screen individually and participants typed in their answers. They had 25 s to

answer each question. No corrective feedback was given. The second initial test was presented

immediately after the first initial test. It contained the same questions but with a new random

presentation order. During Phase 2, participants in the 20 min condition completed the distractor

tasks following the initial test phase, whereas participants in the 24 hr and 7-day conditions

completed the distractor tasks following the appropriate retention interval. In Phase 3,

participants answered 24 questions for each article. Whether participants completed the test for

the tested or control article first was determined randomly. Questions in the tested article were

presented in a random order. However, the nontested questions always appeared *before* the

tested questions to avoid the possibility that output interference may mask the benefits of RIFA

(Roediger, 1974). For the control article, all 24 questions appeared in a random order.[1]

Results

All statistical analyses were carried out with an alpha level of .05. Partial eta squared

(*pes*) indicates effect size for analysis of variance (ANOVA) and Cohen's *d* indicates effect size

for t-tests. Results from the initial test are reported briefly for the sake of completeness, but the

important data are those from the final test.

Initial Test Results

Overall, participants performed better on the second initial test ($M = .63$) than on the first

($M = .60$), $F(1, 81) = 9.68$, *pes* $= .11$, a finding known as hypermnesia (Erdelyi & Becker, 1974).

Not surprisingly, delay condition had no impact on the level of hypermnesia, $F < 1$, nor did it

affect the level of recall on the initial test (averaged initial test performance was .58, .63, and .62

for participants in the 20 min, 24 hr, and 1 week condition, respectively), $F < 1$, which can be

considered evidence against any a priori group differences between participants in the three

delay conditions. Indeed, even if participants in the three delay conditions were matched based

on initial test performance (by selecting the top 24 performing participants in the 20 min

condition), none of the conclusions in the final test changed.[2]

Final Test Results

Figure 1 shows the effects of retention interval on RIFA and the testing effect. The size

of the RIFA effect can be gleaned from a comparison between the middle and bottom lines in

Figure 1, whereas the size of the testing effect is represented by a comparison between the top

and bottom lines. Although neither RIFA nor RIFO was found after 20 min, RIFA was observed

after 24 hr and one week. Second, the effects of initial testing on forgetting differed dramatically

depending on the delay interval (i.e., 20 min to 24 hr vs. 24 hr to 7 days) considered. When one

examines the results between the 20 min and 24 hr interval (see the short segment of lines in

Figure 1), rapid forgetting occurred for the control materials but not for the tested and nontested-

related materials, which suggests that testing slowed forgetting of these materials. However, when one inspects the data between the 24 hr and 7 days interval (see the long segment of lines in Figure 1), all question types demonstrated similar levels of forgetting, which suggests that initial testing had no influence on forgetting. I now back up these impressions with statistical analyses.

Does RIFA Occur After 1 Week? A 2 (question type) X 3 (delay) repeated measures ANOVA revealed a significant main effect of question type, $F(2, 80) = 40.35$, *pes* = .50, a main effect of delay, $F(2, 81) = 9.89$, *pes* = .20, and a significant interaction between question type and delay, $F(4, 162) = 4.25$, *pes* = .10. The main effect of question type suggests that, when ignoring delay, the RIFA effect (7%, $t[83] = 2.94$, $d = .31$) and the testing effect (20%, $t[83] = 8.17$, $d = .93$) were both significant. The main effect of delay suggests that, unsurprisingly, recall probability dropped over time. Before considering the implications of the interaction between delay and question type, it is important to first address the major question of interest. That is, was RIFA found at the 1-week delay? The answer is yes. Indeed, although no RIFA (or RIFO) was found in the 20 min condition (see also, Chan, 2009; Chan et al., 2006), $t < 1$, significant RIFA was observed in both the 24 hr, $t(27) = 2.36$, $d = .46$, and 1-week conditions, $t(27) = 3.00$, $d = .70$ (compare the black squares with the white circles in Figure 1). In addition, a significant testing effect was found for all retention intervals (compare the white triangles to the white circles in Figure 1), all ts > 2.16. Clearly, the benefits of testing on subsequent recall can be long-lasting for both the tested and nontested materials.

What are the Effects of Testing on Forgetting? The interaction between delay and question type can be better understood by an examination of Figure 1. Specifically, at the shorter retention intervals (up to 24 hr, see the short segment of lines in Figure 1), testing exerted a powerful

influence on forgetting; but this influence disappeared when the retention interval was extended

beyond 24 hr (i.e., between 24 hr and 7 days, see the long segment of lines). These impressions

were confirmed by separate ANOVAs examining the data at the shorter (20 min to 24 hr) and

longer (24 hr to 7 days) retention intervals (see Slamecka & Katsaiti, 1988, for a similar

analysis). Between the 20 min and 24 hr interval, question type and delay produced a robust

interaction, $F(2, 53) = 5.52$, $pes = .17$, such that both the tested questions (5% improvement) and

the nontested-related questions (1% forgetting) showed minimal change in recall over the delay,

both ts < 1, but performance on the control questions dropped sharply (12% forgetting), $t(54) =$

2.43, $d = .65$.

A different pattern emerges when one considers the data between the 24 hr and 7 day

interval. The right half of Figure 1 clearly shows that testing had no influence on the rate of

forgetting regardless of question type (see the parallel forgetting lines). This conclusion is

supported by the nonsignificant interaction between question type and delay, $F < 1$, and that all

question types demonstrated similar, and significant, forgetting. For example, the tested items

showed 13% forgetting, $t = 2.23$, $d = .60$, the nontested-related items showed 15% forgetting, $t =$

2.50, $d = .67$, and the control items showed 18% forgetting, $t = 4.12$, $d = 1.12$. The similarity in

the forgetting rates of the different item types is especially noteworthy in light of the fact that

testing has virtually eliminated forgetting for the tested items during the first 24 hours after

initial testing. In fact, based on the data during the first 24 hours, one might conclude that testing

can inoculate against forgetting, but the "normal" forgetting rate beyond the first 24 hours tells a

more complex story.

<center>Discussion</center>

There is a recent surge of interest on the implication of testing on education. This interest stems mainly from the fact that testing has been established as one of the most powerful memory enhancers available (Carpenter et al., 2007; Karpicke & Roediger, 2008; McDaniel, Roediger et al., 2007). Echoing this general idea, recent research (e.g., Chan, 2009; Chan et al., 2006; Cranney et al., 2009; Pilotti et al., 2009) has reported that testing can enhance retention of both the tested and nontested-related materials. In the current experiment, I have demonstrated that the benefits of RIFA are long-lived. Indeed, the benefits of RIFA were equally prominent at the 24 hr and the 7-day intervals. If anything, RIFA increased numerically from 24 hr (9% advantage) to 7 days (13% advantage). Although whether the RIFA effect would exist beyond one week is an empirical question, the fact that RIFA is present after a delay of seven days stands as impressive evidence in favor of applying frequent testing to enhance education. The persistence of RIFA is particularly remarkable when one considers that, unlike the tested materials, the nontested-related materials were virtually never overtly recalled during the initial test. Not only is this finding notable from a pedagogical perspective, it also rules out the possibility that RIFA is based on mechanisms similar to those in producing semantic priming.

As noted in the Introduction, the evidence on whether testing can benefit or harm long-term retrieval of the nontested materials is mixed, such that some studies showed a benefit (Carpenter et al., 2007; Chan, 2009; Chan et al., 2006; Pilotti et al., 2009), some showed a deficit (Garcia-Bajos et al., 2009; Tandoh & Naka, 2007), and still others showed no difference (MacLeod & Macrae, 2001). Although integration has been established as a critical factor in whether testing will help or hurt later recall of the nontested materials (Anderson et al., 2000; Bauml & Hartinger, 2002; Chan, 2009), the effects of delay is still uncertain. Indeed, despite several studies reporting that delay can reduce the negative impact of retrieval inhibition on

recall of the nontested material, a few studies have reported that such inhibition can be long-lasting (Garcia-Bajos et al., 2009; Tandoh & Naka, 2007).[3] Therefore, the current study provided further evidence in this important, and still quite early, line of investigation. Undoubtedly, more research needs to be conducted to further elucidate the theoretical underpinnings of the influence of delay on RIFO and RIFA. What has been shown here is that testing can definitely enhance long-term recall of the nontested materials.

Another major question of the current investigation is: What are the effects of testing on forgetting? Two major viewpoints exist. For example, some researchers believe that testing enhances long-term retention by slowing forgetting (Carpenter et al., 2008; Roediger & Karpicke, 2006b; Runquist, 1983; Wheeler, Ewers, & Buonanno, 2003), whereas others suggest that testing augments initial learning without changing the forgetting function (Pashler, Rohrer, Cepeda, & Carpenter, 2007; Slamecka & Katsaiti, 1988). Here I discuss the current findings in the context of these two viewpoints. I refer to the former hypothesis as the forgetting function argument and the latter hypothesis as the initial learning argument.

Interestingly, the present findings provided support for both arguments. On the one hand, testing flattened the forgetting functions of both the tested and nontested-related materials in the short-term. Specifically, testing practically eliminated forgetting of the tested material and dramatically reduced forgetting of the nontested-related material (relative to the control material) the first 24 hours after encoding. This finding supports the forgetting function argument. On the other hand, the data between the 24 hr and 7-day retention intervals reveal a pattern that is consistent with the initial learning argument. Specifically, participants performed better on the tested and nontested-related questions than on the control questions 24 hours after encoding, but this advantage was *maintained*, *not increased*, during the subsequent six days, such that the drop

in performance for all item types were similar during this delay period. How can one reconcile these seemingly contradictory findings within one experiment?

Although results of the current experiment seem puzzling at first blush, they are not without precedence. Specifically, one interesting pattern has emerged from the few studies that have examined the effect of testing on forgetting over *three or more retention intervals*. Indeed, many studies have shown that testing retards forgetting of the tested material relative to the control material, but this pattern is found only when the origin of the forgetting curve (i.e., the short retention interval) is relatively immediate (Roediger & Karpicke, 2006a; Thompson, Wenger, & Bartling, 1978; Wheeler et al., 2003; Wheeler & Roediger, 1992). When the origin of the forgetting function is at least a day after the initial test (usually in studies in which three or more retention intervals are involved), testing seems to have no influence on forgetting (Roediger & Karpicke, 2006a; Runquist, 1983; Slamecka & Katsaiti, 1988). As can be seen in Figure 2, Runquist (1983), Slamecka and Katsaiti (1988), Roediger and Karpicke (2006b), and Carpenter and colleagues (2008) have all demonstrated patterns similar to the current one, even though they all used different materials and procedures. As a result, the present pattern converges on a growing trend that testing affects forgetting differently depending on the period of retention interval considered. One note of caution here is that different analyses can lead to vastly different conclusions with regard to forgetting (e.g., Bogartz, 1990; Carpenter et al., 2008; Loftus, 1985); therefore, it might be premature to conclude that testing cannot slow forgetting over the long haul. The novel finding here is that recall of the nontested-related materials behaved in a similar manner to the tested materials, such that it too showed attenuated forgetting (compared with the control materials) shortly after the initial test, but this attenuation diminished after 24 hours. From an applied perspective, perhaps the most important finding is that testing

slows forgetting when it occurs the fastest (i.e., soon after learning), and this is an invaluable advantage in educational settings.

Although some exceptions exist (Ballard, 1913; Mitchell, 2006; Wright, Santiago, Sands, Kendrick, & Cook, 1985), the fact that forgetting occurs with delay is virtually inevitable. The study of forgetting has had a long tradition, and researchers have suggested many methods to combat forgetting. The present study shows that testing can drastically reduce forgetting of both the tested and nontested materials, and this advantage is long-lasting.

References

Abbott, E. E. (1909). On the analysis of the factors of recall in the learning process. *Psychological Monographs, 11*, 159-177.

Anderson, M. C. (2003). Rethinking interference theory: Executive control and the mechanisms of forgetting. *Journal of Memory and Language, 49*, 415-445.

Anderson, M. C., Green, C., & McCulloch, K. C. (2000). Similarity and inhibition in long-term memory: Evidence for a two-factor theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 1141-1159.

Ballard, P. B. (1913). Oblivescence and reminiscence. *British Journal of Psychology Monograph Supplements, 1*, 1-82.

Bangert-Drowns, R. L., Kulik, J. A., & Kulik, C. C. (1991). Effects of frequent classroom testing. *Journal of Educational Research, 85*, 89-99.

Bauml, K.-H., & Hartinger, A. (2002). On the role of item similarity in retrieval-induced forgetting. *Memory, 10*, 215-224.

Bjork, E. L., Bjork, R. A., & MacLeod, M. D. (2006). Types and consequences of forgetting: Intended and unintended. In L.-G. Nilsson & N. Ohta (Eds.), *Memory and society: Psychological perspectives* (pp. 134-158). New York: Routledge and Psychology Press.

Bogartz, R. S. (1990). Evaluating forgetting curves psychologically. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*, 138-148.

Burgess, P. W., & Shallice, T. (1996). Confabulation and the control of recollection. *Memory, 4*, 359-411.

Callender, A. A., & McDaniel, M. A. (2007). The benefits of embedded question adjuncts for low and high structure builders. *Journal of Educational Psychology, 99*, 339-348.

Carpenter, S. K., Pashler, H., & Vul, E. (2007). What types of learning are enhanced by a cued recall test? *Psychonomic Bulletin & Review, 13*, 826-830.

Carpenter, S. K., Pashler, H., Wixted, J. T., & Vul, E. (2008). The effects of tests on learning and forgetting. *Memory & Cognition, 36*, 438-448.

Chan, J. C. K. (2009). When does retrieval induce forgetting and when does it induce facilitation? Implications for retrieval inhibition, testing effect, and text processing. *Journal of Memory and Language, 61*, 153-170.

Chan, J. C. K., & McDermott, K. B. (2007). The testing effect in recognition memory: A dual process account. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*, 431-437.

Chan, J. C. K., McDermott, K. B., & Roediger, H. L. (2006). Retrieval-induced facilitation: Initially nontested material can benefit from prior testing of related material. *Journal of Experimental Psychology: General, 135*, 553-571.

Conroy, R., & Salmon, K. (2005). Selective postevent review and children's memory for nonreviewed materials. *Journal of Experimental Child Psychology, 90*, 185-207.

Conroy, R., & Salmon, K. (2006). Talking about parts of a past experience: The impact of discussion style and event structure on memory for discussed and nondiscussed information. *Journal of Experimental Child Psychology, 95*, 278-297.

Conway, A. R. A., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review, 12*, 769-786.

Cranney, J., Ahn, M., McKinnon, R., Morris, S., & Watts, K. (2009). The testing effect, collaborative learning, and retrieval-induced facilitation in a classroom setting. *European Journal of Cognitive Psychology, 21*, 919-940.

Erdelyi, M. H., & Becker, J. (1974). Hypermnesia for pictures: Incremental memory for pictures but not words in multiple recall trials. *Cognitive Psychology, 6*, 159-171.

Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse Processes, 25*, 285-307.

Ford, R. M., Keating, S., & Patel, R. (2004). Retrieval-induced forgetting: A developmental study. *British Journal of Developmental Psychology, 22*, 585--603.

Garcia-Bajos, E., Migueles, M., & Anderson, M. C. (2009). Script knowledge modulates retrieval-induced forgetting for eyewitness events. *Memory, 17*, 92-103.

Hamaker, C. (1986). The effects of adjunct questions on prose learning. *Review of Educational Research, 56*, 212-242.

Karpicke, J. D., & Roediger, H. L. (2007). Expanding retrieval practice promotes short-term retention, but equally spaced retrieval enhances long-term retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*, 704-719.

Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science, 319*, 966-968.

Kornell, N., Hays, M. J., & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*, 989-998.

Leeming, F. C. (2002). The exam-a-day procedure improves performance in psychology classes. *Teaching of Psychology, 29*, 210-212.

Loftus, G. R. (1985). Evaluating forgetting curves. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11*, 397-406.

MacLeod, M. D., & Macrae, C. N. (2001). Gone but not forgotten: The transient nature of retrieval-induced forgetting. *Psychological Science, 12*(2), 148-152.

McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology, 19*, 494-513.

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

Migueles, M., & Garcia-Bajos, E. (2007). Selective retrieval and induced forgetting in eyewitness memory. *Applied Cognitive Psychology, 21*, 1157-1172.

Mitchell, D. B. (2006). Nonconscious priming after 17 years: Invulnerable implicit memory? *Psychological Science, 17*, 925-929.

Pashler, H., Rohrer, D., Cepeda, N. J., & Carpenter, S. K. (2007). Enhancing learning and retarding forgetting: Choices and consequences. *Psychonomic Bulletin & Review, 14*, 187-193.

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*, 179-203.

Roediger, H. L. (1974). Inhibiting effects of recall. *Memory & Cognition, 2*(2), 261-269.

Roediger, H. L., & Karpicke, J. D. (2006a). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17*, 249-255.

Roediger, H. L., & Karpicke, J. D. (2006b). The power of testing memory: Basic research and

      implications for educational practice. *Perspectives on Psychological Science, 1*, 181-210.

Runquist, W. N. (1983). Some effects of remembering on forgetting. *Memory & Cognition, 11*,

      641-650.

Slamecka, N. J., & Katsaiti, L. T. (1988). Normal forgetting of verbal lists as a function of prior

      testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*, 716-

      727.

Spitzer, H. F. (1939). Studies in retention. *Journal of Educational Psychology, 30*, 641-656.

Storm, B. C., Bjork, E. L., Bjork, R. A., & Nestojko, J. F. (2006). Is retrieval success a necessary

      condition for retrieval-induced forgetting? *Psychonomic Bulletin & Review, 13*, 1023-

      1027.

Tandoh, K., & Naka, M. (2007). Durability of retrieval-induced forgetting. *Japanese Journal of

      Psychology, 78*, 310-315.

Thompson, C. P., Wenger, S. K., & Bartling, C. A. (1978). How recall facilitates subsequent

      recall: A reappraisal. *Journal of Experimental Psychology: Human Learning and

      Memory, 4*, 210-221.

Wheeler, M. A., Ewers, M., & Buonanno, J. F. (2003). Different rates of forgetting following

      study versus test trials. *Memory, 11*, 571-580.

Wheeler, M. A., & Roediger, H. L. (1992). Disparate effects of repeated testing: Reconciling

      Ballard's (1913) and Bartlett's (1932) results. *Psychological Science, 3*, 240-245.

Wright, A. A., Santiago, H. C., Sands, S. F., Kendrick, D. F., & Cook, R. G. (1985). Memory

      processing of serial lists by pigeons, monkeys, and people. *Science, 229*, 287-289.

Author Notes

Footnotes

[1] To examine whether output interference affected results of the control questions, they were split into two halves based on presentation order. The recall probabilities for the first and second half questions were .55 and .53 for the 20 min condition, .41 and .43 for the 24 hr condition, and .24 and .24 for the 7-day condition. If output interference was evident, then participants should have performed more poorly on the second half questions than on the first half – No such evidence was found, all ts < 1.

[2] Based on their initial test results, the top 24 performing participants in the 20 min condition yielded an initial test recall rate of .62, which was similar to those in the other delay conditions. Final recall probabilities for these participants were .57, .56, and .65 for the control, nontested-related, and tested items, respectively. Clearly, these final recall probabilities matched up well with those reported in the main text of the manuscript, suggesting that subject selection did not contribute to the findings of the final test.

[3] A few other studies have also reported long-term RIFO (Conroy & Salmon, 2005, 2006; Ford, Keating, & Patel, 2004; Migueles & Garcia-Bajos, 2007; Storm, Bjork, Bjork, & Nestojko, 2006). However, these studies have all used procedures that differed significantly from the typical retrieval practice paradigm. See Chan (2009) for a detailed review of these studies.

Figure Captions

Figure 1. Probability of correct recall as a function of question type (tested, nontested-related, control) and delay (20 min, 24 hr, 7 days). RIFA refers to retrieval-induced facilitation. Error bars are within subjects *.95 CI*

Figure 2. Probability of correct recall as a function of testing condition and delay. The top left panel is based on data from the 9 s encoding condition in Experiment 3 of Runquist (1983). The bottom left panel is based on data averaged across all encoding conditions in Experiment 1of Slamecka and Katsaiti (1988). The top right panel is based on data in Experiment 1 of Roediger and Karpicke (2006b). The bottom right panel is based on data in Experiment 1 of Carpenter et al. (2008); only the data from the 5 min, 1 day, and 7 days delay conditions are displayed here to better match the current experiment and the remaining studies.
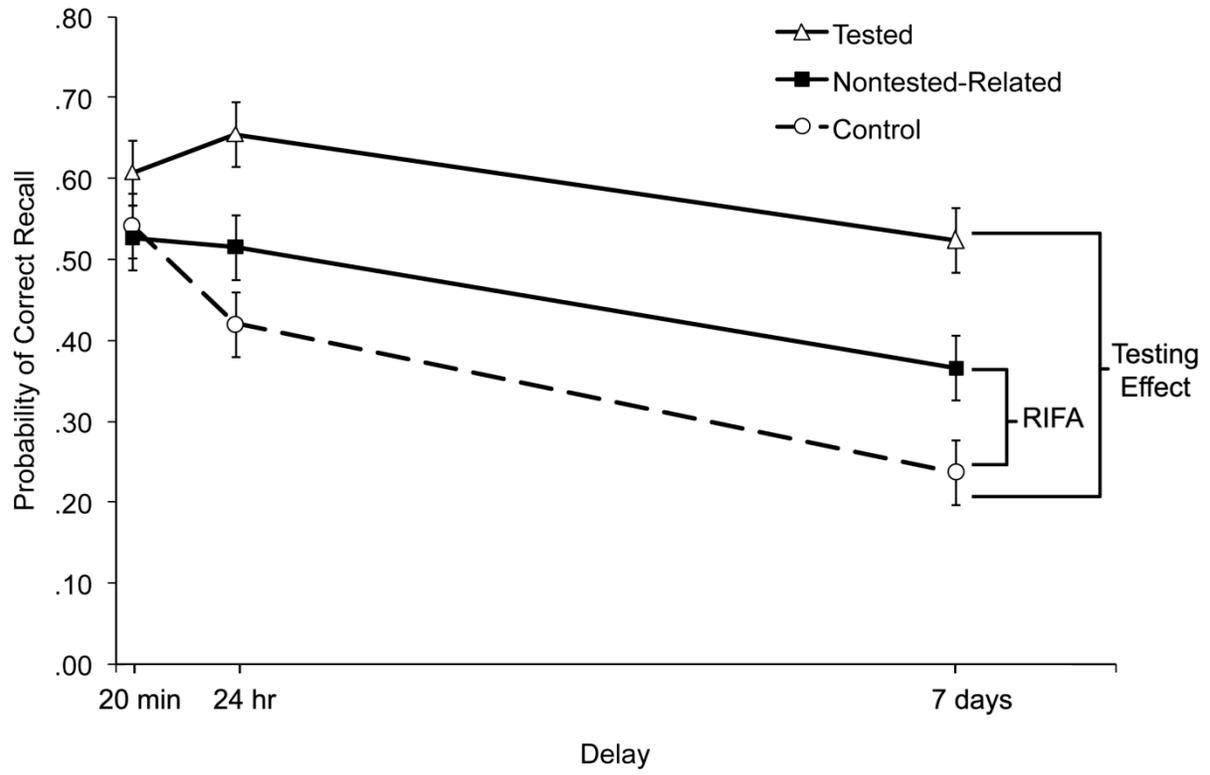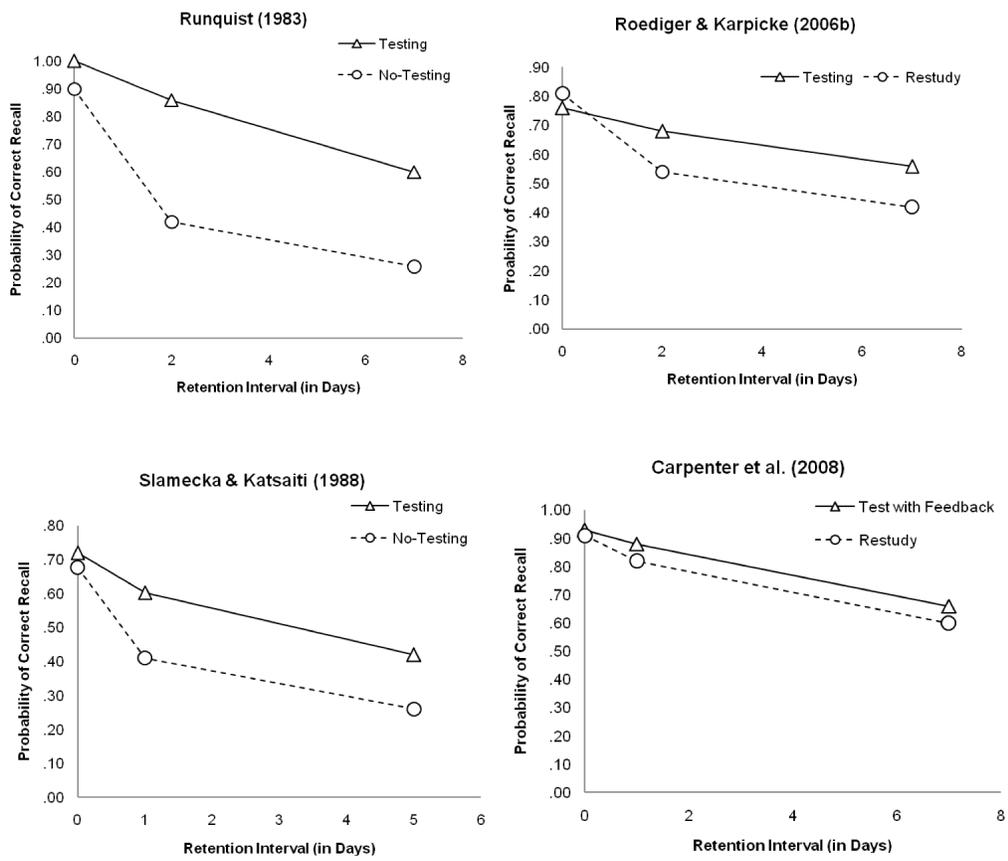
Figure 1.

Figure 2.