Journal of Memory and Language 65 (2011) 85–97

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

Why tests appear to prevent forgetting: A distribution-based bifurcation model

Nate Kornell[a,*], Robert A. Bjork[b], Michael A. Garcia[b]

[a] Department of Psychology, Williams College, Williamstown, MA 01267, USA
[b] Department of Psychology, University of California, Los Angeles, 1285 Franz Hall, Los Angeles, CA 90095-1563, USA

ARTICLE INFO

Article history:
Received 21 September 2010
revision received 6 April 2011
Available online 13 May 2011

Keywords:
Learning
Memory
Testing
Retrieval
Retention interval

ABSTRACT

Retrieving information from memory produces more learning than does being presented with the same information, and the benefits of such retrieval appear to grow as the delay before a final recall test grows longer. Recall tests, however, measure the number of items that are above a recall threshold, not memory strength per se. According to the model proposed in this paper, tests without feedback produce bifurcated item distributions: Retrieved items become stronger, but non-retrieved items remain weak, resulting in a gap between the two classes of items. Restudying items, on the other hand, strengthens all items, though to a lesser degree than does retrieval. These differing outcomes can make tested items appear to be forgotten more slowly than are restudied items—even if all items are forgotten at the same rate—because the test-induced bifurcation leaves items either well above or well below threshold. We review prior evidence and present three new experiments designed to test the bifurcation interpretation.

Introduction

Many college departments use difficult introductory classes as "weeders." The best students learn a lot from these classes, whereas other students look for a new major. If University Tech has weeder Econ and University State does not, Tech will graduate fewer Econ majors, but they will have more extensive training. If all of the Econ majors from both schools were to take the same comprehensive Econ exam upon graduation, the total number passing the exam might be greater at State (given its larger number of Econ majors), but if the same students took the same exam 5 years later, the advantage might be reversed, because many of the more highly selected and well-trained Tech students might still be able to pass the test, whereas many of the State students might fall below the threshold for passing.

The effect of weeder classes on students is analogous to the effect of tests without feedback on learning (e.g., of word pairs). Tests help strong items a lot, but weaker items that cannot be retrieved on the test "drop out." Restudying, by contrast, helps all of the items, but not as much as successful retrieval helps retrieved items. In this article, we argue that one reason why tests appear to prevent subsequent forgetting is because tests bifurcate distributions of to-be-learned items into strong and weak items. As time passes and all items become less accessible, restudied items pass below the final-test threshold before the stronger tested items do so. This bifurcation can produce the appearance of differential forgetting rates even if restudied and tested items are forgotten at the same actual rate as measured by loss of retrieval strength.

The difference between recall and memory strength

Psychologists often measure learning using cued-recall or free-recall tests. Recall tests produce one of two outcomes—the correct answer is recalled or it is not. Models

* Corresponding author. Fax: +1 413 597 2085.
E-mail addresses: nkornell@gmail.com (N. Kornell), rabjork@psych.ucla.edu (R.A. Bjork), gikeymarcia@gmail.com (M.A. Garcia).

0749-596X/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jml.2011.04.002

of memory tend to assume, however, that items vary in a continuous way along one or more dimensions of strength (see, e.g., Bjork & Bjork, 1992). Such strengths may not be measurable directly, and may not be meaningful in isolation—because retrieval success depends on what cues are available at the time of retrieval, whether to-be-recalled items were recently primed, and so forth—but all else being equal, models assume that items with more strength are more retrievable. For current purposes, we assume that (a) memory strengths of different items lie on a continuum and (b) an item can be recalled if its current strength in memory is above some recall threshold.

Given those assumptions, it is important to emphasize that recall tests do not measure an item's memory strength directly. Instead, they measure whether that item's strength is high enough to surpass a threshold for recall. As a result, the set of memories with the greatest total strength is not necessarily the set of memories that produces the highest average recall accuracy. To illustrate, assume we have conducted an experiment with 10 items in each of two conditions, where the memory strengths in Condition 1 are 60, 60, 60, 60, 60, 0, 0, 0, 0, and 0 on an arbitrary 0–100 scale, whereas the strengths in Condition 2 are 100, 100, 100, 100, 40, 40, 40, 40, 40, 40. If the threshold for recall is 50, accuracy in Condition 2 will be 40% (4 out of 10), whereas the accuracy in Condition 1 will be 50% (5 out of 10). Thus, average accuracy is higher in Condition 1 than in Condition 2, even though the average memory strength in Condition 1 (30) is less than half the average memory strength in Condition 2 (64).

In the example above, average memory strength and recall accuracy diverged because we assumed quite bizarre distributions of memory strengths across items. It is probably safe to assume that under most circumstances, memory strengths are roughly normally distributed. Testing, however, bifurcates and distorts such distributions, as we argue in more detail below.

Testing effects

Testing-effect experiments often include three crucial conditions: a restudy condition, a test condition, and a control condition. In the *restudy* condition, participants study items and then restudy those same items at a later time, usually by re-reading those items. In the *test* condition, they study the items and then take a test on those items at a later time. In the *control* condition, the items are studied, but then are not restudied or tested before some final criterion test. The results of many different experiments (for a comprehensive review, see Roediger & Karpicke, 2006b) have demonstrated that an initial test not only enhances recall on a later test—as measured against such a control condition—but often also yields better recall on the final test than does the restudy condition. (Note that although a test is a way to restudy, we use the term *restudy* in this paper to refer to restudy trials that do not involve a test.)

One of the most intriguing effects of tests is that the benefits of tests appear to grow more pronounced as the retention interval from restudying or testing to a final test increases. That is, there is a *test-delay interaction*. For example, what appears to be an advantage of restudying over testing on an immediate test can change into an advantage of testing over restudying on a delayed test (e.g., Roediger & Karpicke, 2006a). Later in this article we review research demonstrating the test-delay interaction.

In this paper, our goal is not to explain the mechanisms underlying the testing effect. We simply assume, based on a plethora of prior evidence, that a successful retrieval enhances learning more than does a presentation of the same information (see Roediger & Karpicke, 2006b, for a review). Our goal, drawing on the distribution framework, is to clarify why tests appear to prevent forgetting.

Why tests appear to prevent forgetting

The current paper was motivated by the question: How shall we explain the test-delay interaction? One explanation is that retrieving information has specific effects on forgetting. In other words, tests do not just enhance learning, they also prevent forgetting. If true, this property of retrieval would explain the test-delay interaction.

We suggest an alternate explanation of the test-delay interaction. It is important to acknowledge at the outset that the explanations are not mutually exclusive and it is possible that both contribute to the test-delay interaction. The explanation offered here has evolved over decades of thinking by the second author (Bjork, Hofacker, & Burns, 1981; Gelfand, Bjork, & Kovacs, 1983; Halamish & Bjork, in press).

To understand this alternate explanation, it is necessary to return to the idea that tests can produce non-normal distributions of memory strengths across items. Consider first what happens when learners are not shown the correct answer—that is, they are not given feedback—after the test trials. The no-feedback methodology has been used in most or all experiments demonstrating the test-delay interaction, as we explain below.

Assume there are two sets of items, set A and set B, and that items in set A are tested without feedback and the items in set B are re-presented for study. For the items in set A that are successfully retrieved, the act of retrieval serves as a potent learning event, and the memory strengths of these items are boosted by a large amount. Items that are not retrieved are not boosted at all (see Spellman & Bjork, 1992), at least under ordinary circumstances. Moreover, they are already below threshold, by virtue of not being recallable, and they gradually drift toward being less and less memorable owing to the inexorable process of forgetting. Thus, there are two classes of items in item set A: items that get a big memory boost because they are retrieved and items that are not boosted at all. That is, tests bifurcate item distributions. The items in set B, on the other hand, are all presented for restudy, which means that they all get a boost, but not as large a boost as the retrieved items in set A. They end up normally distributed, with the resulting distribution falling between the two sections of the test-bifurcated distribution in set A.

The resultant memory strengths are illustrated in Fig. 1. Before restudying or testing, the two sets of items are

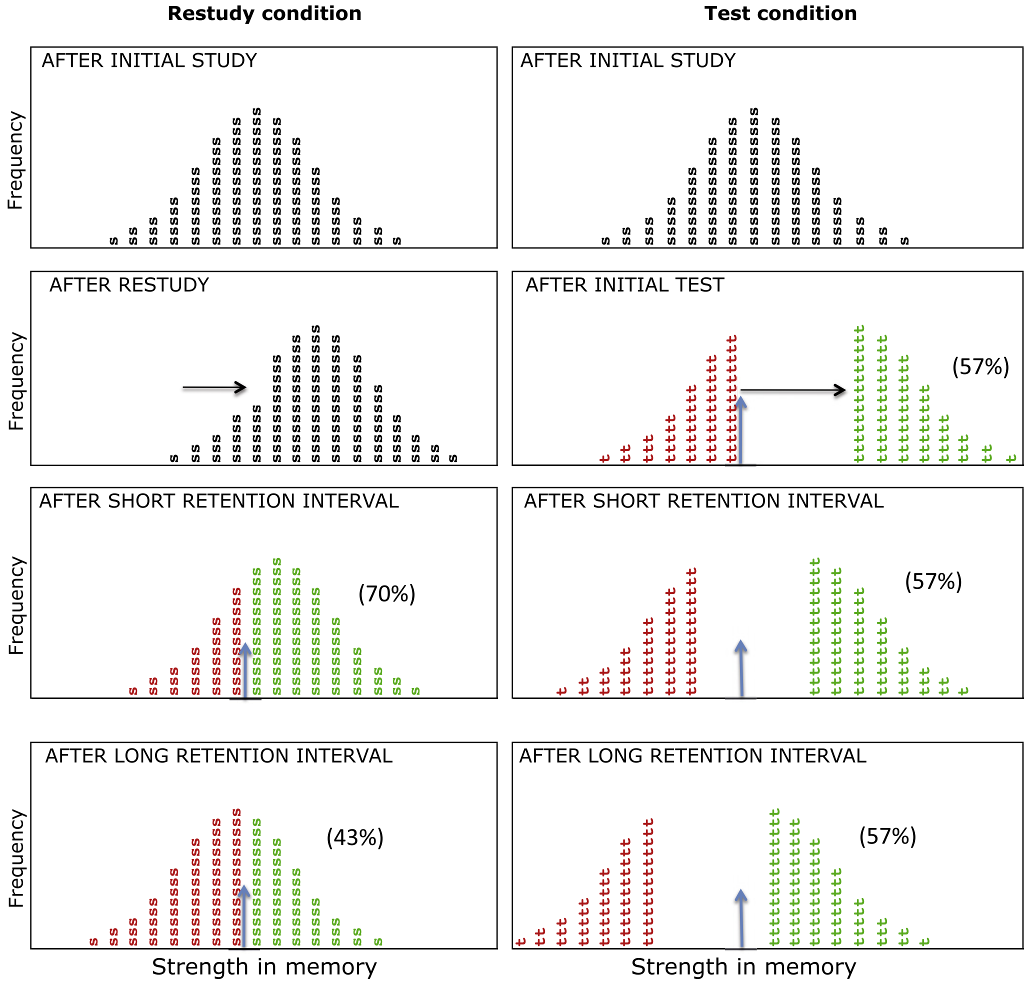

Fig. 1. Simulated memory strength for two hypothetical sets of 100 items. The left and right column represent items that were restudied or tested without feedback, respectively. The top panels represent memory strength after initial study. The two distributions are identical and both are normally distributed. In the second pair of panels, the restudy items (left panel) all gain memory strength equally; the tested items (right panel) become bifurcated. Items that are above threshold (i.e., that are retrieved) gain more strength than the restudied items gain, whereas items below threshold do not gain memory strength. The vertical arrow represents the recall threshold. The next pair of panels represent memory strength after a short retention interval; the bottom pair of panels represent memory strength after a long retention interval. All items are forgotten (i.e., move leftward) at the same rate, but the bifurcated distribution in the test condition appears to prevent forgetting when measured by the percentage of items that are recallable (i.e., above threshold).

normally distributed with the same mean and standard deviation. After the items are restudied (left column) or tested (right column) the resulting distributions differ markedly. Subsequently, if recall is measured after a short retention interval, restudying appears to be advantageous, whereas after a long retention interval, testing appears to be more effective (in both cases, as measured by the proportion of items above the presumed recall threshold indicated by the vertical arrow).

The result of forgetting, given the distributions in Fig. 1, is that recall accuracy shows the test-delay interaction. Notice, though, that the forgetting rates for the two conditions are *exactly the same*. Thus, the test-delay interaction can be explained even if it is not tests, but rather the differential strengthening a subset of items, that appears to prevent forgetting. We will refer to this explanation of the test-delay interaction as the *bifurcation model*, because it rests on the assumption that a test bifurcates the distribution of item strengths.

A simple model of memory

As stated earlier, our goal is not to provide an explanation the mechanisms underlying the testing effect. Fig. 1 is a simplified (even crude) model that does not reflect the complicated dynamics of learning and forgetting. This model is conservative, and perhaps biased against our hypothesis, in at least two ways. First, we assume that all items are forgotten at the same rate, but well-known information may be forgotten more slowly than less well-known information (Bjork & Bjork, 1992). As a result, tested items may be forgotten more slowly than restudied items, but not because of the benefits of testing per se. Instead, it may be that tests, like other manipulations (e.g., deep processing; Craik & Tulving, 1975), create strong memories, and strong memories last. Second, forgetting curves are, at least as measured by recall tests, negatively accelerated.

For both of these reasons, the strongest items should take even longer, relative to weaker items, to cross below threshold than our model predicts. Thus, our model may underestimate the explanatory power of the bifurcation model. We contend, however, that the simplified model presented in Fig. 1 and a more complex model produce similar changes in apparent forgetting rates as a result of bifurcation. The experiments presented below test the power of this simple model to explain actual recall performance.

Two explanations of the test-delay interaction

The bifurcation model is one way to explain the test-delay interaction. But it is not the only possible explanation. Alternatively, tests could prevent forgetting because the act of retrieval might have specific effects on subsequent forgetting rates. We will refer to this idea as the *retrieval-prevents-forgetting* hypothesis. It entails that there is something special about retrieval and that retrieval affects forgetting in a way that is independent of other measures of memory strength. (A note on terminology: The test-delay interaction refers to a pattern of data, whereas the retrieval-prevents-forgetting hypothesis is a potential explanation of *why* there is a test-delay interaction.)

The two explanations of the test-delay interaction are not mutually exclusive, and it is possible that both coexist. Our primary claims are that (a) the bifurcation model helps explain apparent differences in forgetting rates, and (b) in the absence of direct evidence, it is not necessary to assume that retrieval prevents forgetting.

The role of feedback

Thus far we have considered situations in which test trials are not followed by feedback (i.e., the correct answer is not shown). Feedback may not affect successfully retrieved items (e.g., Pashler, Cepeda, Wixted, & Rohrer, 2005). But feedback can have a large effect on non-retrieved items: Instead of essentially dropping out, these items gain strength. Indeed, evidence reviewed below suggests that unsuccessfully attempting to retrieve an item, and then receiving feedback, results in more learning than does restudying the same item without making a retrieval attempt (e.g., Kornell, Hays, & Bjork, 2009). If non-retrieved items benefit sufficiently from feedback, then feedback may reduce or prevent bifurcation.

Thus, the bifurcation model predicts that providing feedback should, at a minimum, reduce the test-delay interaction. Whether feedback should *eliminate* the test-delay interaction is an open question. If unsuccessful retrieval attempts followed by feedback result in just as much learning as do successful retrieval attempts, there should be no bifurcation following test trials with feedback. If there is something special about retrieving an answer oneself, rather than simply attempting to retrieve that item and then receiving feedback, then there should be some bifurcation between retrieved and unretrieved items, even when feedback is provided.

Unlike the bifurcation model, the retrieval-prevents-forgetting hypothesis does not predict that feedback should reduce the test-delay interaction. It predicts that retrieval should prevent forgetting even if the retrieval attempts are followed by feedback. We test these competing hypotheses in Experiment 2.

Empirical support for the test-delay interaction

Roediger and Karpicke's (2006b) excellent review of nearly a century of testing effect research covers many studies that have shown the test-delay interaction. We review below only a representative sample of these studies. To foreshadow, we conclude (a) that many experiments have shown the test-delay interaction, in both free and cued-recall, and (b) that there appears to be a consensus among researchers that tests slow the rate of forgetting—that is, most researchers appear to endorse the retrieval-prevents-forgetting hypothesis.

Free recall

One of the first studies to show the test-delay interaction was conducted by Hogan and Kintsch (1971). In their first experiment, they asked participants to study 40 words and take free recall or recognition tests. Conditions 1 and 2 involved studying three times and then taking a test (SSST) or studying once and then taking three tests (STTT). Participants recalled more items on the test at the end of the SSST condition (39%) than the STTT condition (30%). Yet a week later, recall performance was the same in the two conditions (20%). Thus, these data show a test-delay interaction.

Hogan and Kintsch's (1971) finding of a test-delay interaction was subsequently replicated by Thompson, Wenger, and Bartling (1978) and Wheeler, Ewers, and Buonanno (2003). The authors of the latter article concluded "that study and test trials have different effects upon memory, with study trials promoting memory acquisition, and test trials enhancing the retrieval process itself, which protects against subsequent forgetting" (p. 571). Based on their evidence, they argued that models of forgetting need to account for differences in forgetting rates that depend on how items are practiced.

Other kinds of free recall task have shown similar results. For example, Roediger and Karpicke (2006a) uncovered one of the most dramatic demonstrations of the test-delay interaction. They asked participants to study short passages on sea otters or the sun. In their second experiment, they found that in the SSSS condition recall dropped from over 80% correct 5 min after participants studied to 40% a week later. In the STTT condition, the drop was from 70% to 60%. The sheer size of this interaction is impressive; as a proportion of the information participants recalled after 5 min, participants in the delayed SSSS conditions forgot 52% of the information, whereas participants in the STTT condition forgot only 14%. The authors conclude that the data "clearly demonstrate the powerful effect of repeated testing in preventing forgetting." (p. 253).

Cued-recall

The test-delay interaction has also been shown in cued-recall experiments. For example, Runquist (1983) asked participants in his second experiment to study 24 word pairs and then tested them initially on 12 of the pairs. He then gave them a criterion test either immediately or after a delay. Recall accuracy declined more, as a function of delay, for items that were not tested than it did for items that were tested. He concluded that "the difference between tested and untested items increased with longer retention intervals" (p. 641). Runquist (1986) found similar outcomes, concluding that the results "point to the importance of retrieval operations in the attenuation of forgetting" (p. 65). As Carpenter, Pashler, Wixted, and Vul (2008) point out, however, Runquist (1983, 1986) only analyzed items from the test condition that had been answered correctly during the study phase, whereas he analyzed all items from the restudy condition, introducing a item-selection effect that could explain the benefit of testing in his experiments, although not necessarily the test-delay interaction.

Slamecka and Katsaiti (1988) found evidence that tests slowed the rate of forgetting when they compared 0- and 1-day retention intervals, but they found no difference in forgetting rates between 1- and 5-day retention intervals. Based on this outcome, they concluded that the apparent difference in forgetting rates was illusory, because it did not hold up beyond 24 h delay. Clearly, though, from 0 to 5 days there was a difference in apparent forgetting. They also found, in their third experiment, that when they gave feedback (in the form of a subsequent presentation) the difference in forgetting rates disappeared, even between day 0 and day 1, which is consistent with the hypothesis of the bifurcation model that feedback diminishes or eliminates the test-delay interaction.

Other cued-recall studies have shown similar results. For example, Cull (2000) concluded that in his data, "the benefits of testing were amplified as the delay between the last review and the final test increased." (p. 231). Carpenter et al. (2008), who analyzed forgetting rates via curve fitting as well as ANOVA, concluded that "…in two out of three experiments, testing also reduced forgetting more than restudying, though this was not always the case according to the ANOVA." (Carpenter et al., 2008; also see Pashler, Rohrer, Cepeda, & Carpenter, 2007).

One recent study, by Toppino and Cohen (2009), explicitly set out to investigate "the testing effect and especially its critical interaction with retention interval while eliminating or minimizing methodological concerns that have raised questions about previous experiments." (p. 254). In Experiment 2, Toppino and Cohen provided extensive training "to insure a high level of recall in the testing condition." (p. 254). They compared two conditions, SSSSSSSSS versus SSSSSSSST. Participants recalled 85% of the word pairs on the test during the study phase. Participants who took a final test after a 5-min delay recalled between 85% and 90% of items in both conditions. When the final test was delayed by 48 h, however, more of the studied items than tested items had passed below threshold—that is, there was an unmistakable test-delay interaction.

Toppino and Cohen (2009) did not discuss the effect of item distributions on apparent forgetting, but they may have played an important role. If almost all tested items are retrieved successfully during the study phase (and 85% were recalled in Experiment 2), there will be some bifurcation, but it will probably have small effects. Even in the absence of bifurcation, though, item distributions can explain interactions when recall rates are very high. If almost all items are above threshold in both conditions, and the tested items are farther above threshold than are the studied items, then, as time passes, studied items should begin to cross below threshold before the tested items do so. Thus, even if forgetting rates were the same, differences in item distributions could produce a test-delay interaction. Moreover, adding a ninth study trial after a series of eight study trials probably had little marginal value in Toppino and Cohen's experiment. But adding a test on trial nine may have benefited learning substantially (in part because studying an item one can already recall has little value, but being tested on the same items is very valuable; Karpicke & Roediger, 2007, 2008). Thus, at the end of the study phase in Toppino and Cohen's experiment, the distribution of tested items was probably well above the distribution of studied items. In short, when performance levels are high—and thus, most items start out above threshold—and the distribution of tested items is significantly stronger than the distribution of studied items, a test-delay interaction can be expected even if the bifurcation of the distribution is relatively minimal.[1]

Other studies have, as in Toppino and Cohen's (2009) study, produced very high, or even nearly perfect, levels of recall accuracy during learning (e.g., Karpicke, 2009; Karpicke & Roediger, 2007, 2008; Tulving, 1967). In these studies, participants were typically trained via multiple cycles of studying and/or testing. The initial test cycles typically show less than perfect performance. When there are multiple tests per item, some items will be recalled successfully multiple times, creating high levels of memory strength, whereas others might only be recalled once. For example, if items in subset A were retrieved successfully four times and items in subset B were retrieved only once, the result could be a bifurcated distribution, with A items far stronger than B items, even though all items were eventually recalled. Thus, testing can produce bifurcation even if eventual performance is perfect.

The point here is not to provide an exhaustive review of evidence for the test-delay interaction. Instead, the point is that many studies have shown the test-delay interaction in both cued- and free-recall.

There is another important point. According to the bifurcation model, the test-delay interaction should appear

[1] In the experiments reported herein, we use the logic that answering all items correctly should eliminate the test-delay interaction. Our experiments differ from those of Toppino and Cohen (2009), however, in that performance levels were not near ceiling and we showed that the test-delay interaction diminished when recall success was (virtually) guaranteed.

primarily in the absence of feedback. Feedback was absent in all of the studies described in this section (with the exception of Slamecka & Katsaiti, 1988, as noted above, and Carpenter et al., 2008, who found a small effect in some analyses; we return to the latter findings in the general discussion). Thus, the extant literature appears to be compatible with the bifurcation model.

Novel predictions

The bifurcation model is consistent with previous demonstrations of the test-delay interaction. To further test the model, we investigated two novel predictions. First, according to the model, apparent differences in forgetting happen because some tested items are recalled and some are not. Thus, the bifurcation model predicts that when all tested items are treated equally, the test-delay interaction should diminish or disappear. Second, tests are not the only way distributions can be bifurcated. The bifurcation model predicts that bifurcating a distribution should appear to prevent forgetting even when no items are tested.

Why treating all items equally eliminates the test-delay interaction

The test-delay interaction occurs when retrieved items are boosted while unretrieved items are not. According to the bifurcation model, treating all items equally should either diminish or eliminate the test-delay interaction. (Although as we argued above, if almost all tested and untested items are above threshold, there could be an interaction due to distributional differences even if the role of bifurcation per se is minimal.) The tests-prevent-forgetting hypothesis predicts that treating all items equally should not prevent a robust test-delay interaction from occurring.

One way to treat all items equally is to ensure that no item is retrieved successfully. Because no items are retrieved, no subset of items is systematically boosted more than any other subset. Kornell et al. (2009) conducted a series of experiments that did just that. They compared a test condition to a study condition, but there was no initial study phase, so the participants never answered correctly during the test phase (but they were given feedback). Kornell et al.'s evidence suggested that retrieval attempts enhanced subsequent encoding, even when the retrieval attempts failed (also see Izawa, 1970; Karpicke, 2009; Karpicke & Roediger, 2007; Richland, Kornell, & Kao, 2009).

Kornell et al. (2009) did not find a test-delay interaction. The bifurcation model correctly predicts this outcome: Because all items were treated equally in the test condition, the tested and untested items should have been forgotten at the roughly the same apparent rate. The retrieval-prevents-forgetting hypothesis predicts that the benefits of retrieval, which Kornell et al. found, should grow larger as the retention interval increases. This prediction was not supported.

Some might find reasons to argue, however, that although successful retrieval enhances learning and prevents forgetting, unsuccessful retrieval enhances learning but does not prevent forgetting. Thus, we conducted a new experiment. Instead of insuring that no item was retrieved successfully, we insured that *all* items were retrieved successfully.

Experiment 1

In Experiment 1, after studying a set of word pairs once each, participants studied the pairs again in three within-participant conditions. In the Test condition, the cue was presented, and participants were asked to recall the target and type it in. In the Consonants condition, the cue was presented along with the target, but the target was missing vowels (e.g., EARTH – PL_N_T), and participants were asked to type in the target (PLANET). In the Copy condition, the cue and target were presented together, intact, and participants were asked to type in the target.

The Consonants condition allows for retrieval of two kinds. First, even if the pairs had not been studied previously, figuring out the answer involves recalling the association (e.g., between planet and earth) from semantic memory. Second, because the items were studied previously, pairs could be recalled based on the study phase in the Consonants condition (as they were in the Test condition). It is true that retrieval with consonants differed in some ways from retrieval without consonants, but it seems to us that the simplest explanation of the Consonant condition's benefits (described shortly) is that they were due to retrieval. (We also address this issue by providing Feedback in Experiment 2.)

We predicted that the apparent rate of forgetting would be greater in the Copy condition than in the Test condition, replicating the test-delay interaction. The interesting question has to do with the rate of forgetting in the Consonants condition. Would it look more like the Test condition or the Restudy condition? Again, there were two competing predictions. If retrieval prevents forgetting, the rate of forgetting in the Consonants condition should be similar to the rate of forgetting in the Test condition, because both involve retrieval. But we expected the rate of retrieval accuracy in the Consonants condition to be near perfect during the study phase, which would mean all items would be treated equally. Thus, according to the bifurcation model, the rate of forgetting in the Consonants condition and the Restudy condition should be about the same—and they should both differ from the Test condition.

Method

Participants

Twenty-two adults, aged 22–51 (mean = 32.5), were recruited using Amazon's Mechanical Turk, a website where people can sign up to complete jobs online. Nine participants were female and 13 male. Ten lived in the United States of America, eight lived in India, and the other four were from four different countries. All participants reported speaking English fluently. They were paid $2.50 for participating in the first session of the experiment, which took about 25 min. They received an additional $2.00 upon completing the second session, which took about 5 min.

Materials

The stimuli were 60 word pairs with forward association strengths of between .15 and .20 according to Nelson, McEvoy, and Schreiber's (1998) norms. That is, when people were shown the first word in the pair, 15–20% of those individuals produced the second word in the pair as the first associate that came to mind. From among the set of words that fit that description, a subset was selected for which it seemed relatively easy to generate the second word in the pair, when presented with the first word intact plus the second word missing vowels. No pair had a second word that began with a vowel. The pairs are included in the appendix.

Design

The experiment employed a 3 × 2 within-participants design. Study Condition had three levels: Test, Consonants, and Copy. Retention Interval had two levels; half of the items were tested after 2 min and half were tested after 2 days.

Procedure

The experiment was conducted online. Following the instructions, there were four phases to the experiment: initial study, practice, immediate test, and delayed test.

During the initial study phase, each pair was presented once. On each trial, the cue and target were presented. Below them, the cue was presented again and participants were asked to copy the target into a box next to the cue. Each trial lasted 5 s.

During the practice phase, all pairs were presented again, with 20 pairs in each of the three conditions. In the Test condition, the cue was presented and participants were asked to type the target in the box next to it. In the Consonants condition, the cue was presented with the target next to it, but all vowels in the target were replaced with underscore characters (e.g., MESS-CL__N). Below this line, the cue was presented again and participants were asked to type the intact target (e.g., CLEAN) into the box. The Copy condition was the same as the initial study phase: The cue and target were presented and the participant was asked to type in the target. Participants had 5 s to answer in all conditions. Feedback was not given in any condition.

The practice phase was repeated twice, in a different random order each time, so that there were 120 trials, with each item being studied twice. (Note that the more test trials there are, according to the bifurcation model, the more the distribution will become bifurcated; thus, the difference in forgetting should become greater, everything else being equal, when comparing STT to SSS than it is when comparing ST to SS. Thus, we included two trials in the practice phase.)

Following the practice phase, there was a 2-min distractor task during which participants were asked to type in the names of as many countries as they could think of.

The first session ended with a test, on which half of the items in each condition (i.e., 10 items from each condition) were tested. During the test the cue was presented and participants were asked to type in the target. They had unlimited time to do so.

Participants were emailed 48 h after the first session and asked to participate in a second session. The test was the same as the test during the first session, in that the cue was presented and participants had as much time as they wanted to try to type in the target. All 60 items from session 1 were tested during session 2. The data analysis was limited to items that had not been tested during session 1.

Results

In the analyses presented herein, we analyzed forgetting rates using interaction terms computed based on ANOVAS, and we compared individual forgetting rates by subtracting delayed recall from immediate recall. It is worth noting that there are other ways of conceptualizing forgetting (e.g., as the proportion of items that were recalled on the immediate test that were lost by the time of the delayed test). We chose to use the subtraction/ANOVA method for two reasons; first, it is the measure of forgetting used most often in prior research. Second, it corresponds well to the idea, built into the bifurcation model, that it is important to consider the number of items above and below threshold.

Participants were asked to type in each answer twice during the practice phase. During the first cycle through the practice phase, recall accuracy averaged .98, .89, and .54 in the Copy, Consonants, and Test conditions, respectively. During the second cycle, the averages were .97, .94, and .55, respectively. (Note that participants running out of time and/or misspelling answers can account for some of the errors.)

Final test performance is displayed in Fig. 2. A 3 × 2 within-participants ANOVA identified significant effects of Study Condition ($F(2, 42) = 24.60$, $p < .0001$, $\eta_p^2 = .54$) and Retention Interval ($F(1, 21) = 25.35$, $p < .0001$, $\eta_p^2 = .55$), and there was a significant interaction ($F(2, 42) = 10.22$, $p < .001$, $\eta_p^2 = .33$).

The key question was how the forgetting rates in the three conditions would compare to each other. Forgetting rates were computed by subtracting recall accuracy in

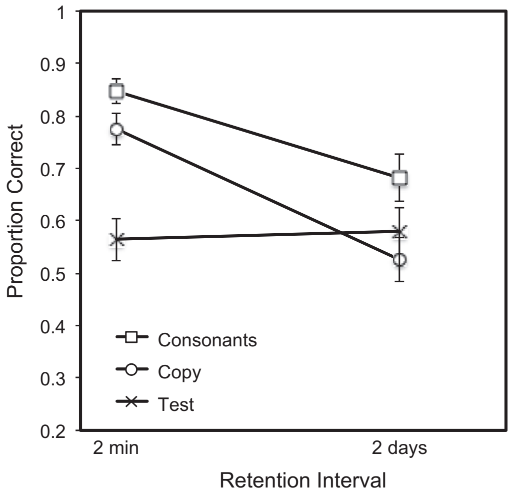

Fig. 2. Proportion correct in Experiment 1. Error bars represent ± 1 standard error of the mean.

the Delayed condition from recall accuracy in the Immediate condition. The forgetting rates were .25, .17, and −.02 in the Copy, Consonants, and Test conditions, respectively. An ANOVA was conducted on the three forgetting rates. The ANOVA was identical to the interaction above, but it allowed for a simple effects comparison using a Tukey test. This comparison showed that there was significantly less observed forgetting in the Test condition than in the Copy and Consonants conditions. The latter conditions did not differ significantly.

Discussion

As the bifurcation model predicted, the forgetting rate in the Consonants condition was similar to the forgetting rate in the Copy condition. This outcome was inconsistent with the prediction of the tests-prevent-forgetting effect hypothesis.

The apparent forgetting rate was smaller in the Consonants condition (17 percentage points) than it was in the Copy condition (25 percentage points). Although this difference was not significant, one way of interpreting the apparent difference is that retrieval does not only bifurcate distributions, it also prevents forgetting, and thus the Consonant condition prevented forgetting by a small amount. This interpretation would support the theory that retrieval prevents forgetting. An alternate interpretation is that some amount of bifurcation did occur in the Consonant condition, where accuracy was not perfect (.89 and .94 during the two practice phases), and this bifurcation contributed to the non-significant difference in apparent forgetting.

It is also worth mentioning that performance in the Test condition did not decrease over the course of 48 h. In fact, it increased, although by a tiny amount and not significantly ($t(21) = -.36$, $p = .72$). We return to this point below.

Experiment 2

In Experiment 2, we used the same three conditions as Experiment 1, but added a fourth condition, the Test-plus-feedback condition, in which test trials were followed by feedback. As explained above, the bifurcation model predicts that providing feedback should lessen or eliminate the bifurcation that occurs due to tests, and, thus, the Test-plus-feedback condition should not appear to prevent forgetting. Again, the tests-prevent-forgetting hypothesis predicts the opposite: A test should prevent forgetting regardless of whether or not it is followed by feedback.

The first goal of Experiment 2 was to test the effects of feedback on the test-delay interaction. The second goal was to replicate and extend Experiment 1. An additional benefit of the Test-plus-feedback condition is that it did not rely on the assumption that the Consonants condition involves retrieval.

Method

The participants were 40 adults recruited using Amazon's Mechanical Turk. They ranged in age from 18 to 61 (mean = 29.3). Nineteen participants were female and 21 male. Twenty-two lived in the United States of America, 11 lived in India, two lived in Canada, and the other five were from five different countries. All participants reported speaking English fluently. They were paid $2.50 for participating in the first session of the experiment, which took about 25 min. They received an additional $2.00 upon completing the second session, which took about 5 min.

The experiment involved a 4 × 2 within-participants design; there were four study conditions (Copy, Consonants, Test, and the new Test-plus-feedback condition). The two retention intervals, 2 min and 2 days, were the same as in Experiment 1. The stimuli were 48 of the 60 word pairs used in Experiment 1.

The procedure of Experiment 2 was similar to the procedure of Experiment 1. One important difference was that feedback was given, for 2 s per trial, during the initial study phase and practice phase. The cue and target were shown together as feedback during the initial study phase. The same was true during the practice phase in the Copy, Consonants, and Test-plus-feedback conditions; in the Test condition, a screen that did not display the cue and target was shown for 2 s. Thus, participants had 5 s to type in their response on all trials followed by 2 s of feedback (or a lack thereof). In Experiment 1, all trials during the initial study and practice phases lasted 5 s and there was never feedback on any trial. The number of items was reduced from 60 to 48 so that the first session of the experiment still took roughly 25 min. Thus, six items were assigned to each of the eight conditions.

Results

During the first cycle through the practice phase, recall accuracy averaged .98, .88, .50, and .48 in the Copy, Consonants, Test-plus-feedback, and Test conditions, respectively. During the second cycle, the averages were .98, .94, .78, and .50, respectively.

The final test results are displayed in Fig. 3. As in Experiment 1, there was a significant main effect of Study Condition ($F(3, 117) = 40.53$, $p < .0001$, $\eta_p^2 = .51$) and Retention Interval ($F(1, 39) = 43.22$, $p < .0001$, $\eta_p^2 = .53$), and there was a significant interaction ($F(3, 117) = 3.79$, $p < .05$, $\eta_p^2 = .09$).

Again, the crucial question has to do with forgetting rates. The forgetting rates in the Copy, Consonants, Test-plus-feedback, and Test conditions, respectively, were .25, .18, .20, and .08. The finding that the forgetting rate in the Test condition was much lower than it was in any of the other conditions replicates Experiment 1. Moreover, the apparent forgetting rate was significantly greater in the Test-plus-feedback condition than it was in the Test condition, $t(39) = 2.36$, $p < .05$.

Discussion

The Test-plus-feedback condition involved the same sort of retrieval as did the Test condition. If retrieval prevents forgetting, it should do so when feedback is provided as well as when it is not. This prediction was not supported

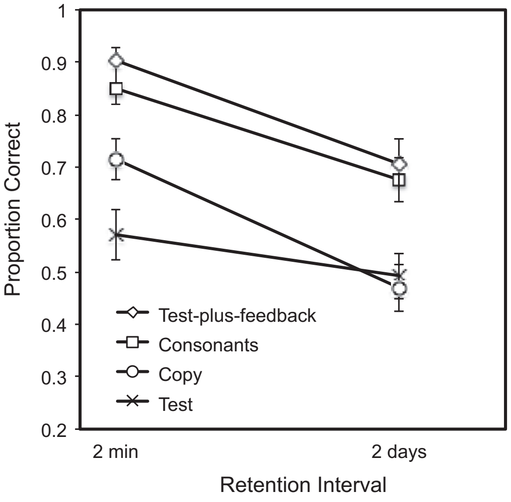

Fig. 3. Proportion correct in Experiment 2. Error bars represent ± 1 standard error of the mean.

in Experiment 2. Instead, items in the Test-plus-feedback, Consonants, and Copy conditions were all forgotten at similar rates. Test items were forgotten at a lower apparent rate. Moreover, there was a significant difference in apparent forgetting rate between the Test-plus-feedback and Test conditions. The data seem inconsistent with the idea that retrieval per se prevents forgetting. They are consistent with the bifurcation model.

Participants in Experiment 1 recalled more items from the Test condition after a delay than they did immediately. Although this finding was intriguing, it was not significant, and Experiment 2 showed that it was not replicable.

Experiment 3

Before presenting the first experiment, we made two predictions based on the bifurcation model. The first prediction was that if all items are treated equally, the test-delay interaction should diminish or disappear. Previous research (Kornell et al., 2009) and Experiments 1 and 2 supported this prediction. The second prediction was that bifurcating a distribution can appear to prevent forgetting even if no items are tested. We turn to this prediction in Experiment 3.

In Experiment 3 we attempted to simulate the kind of distribution that might be created in a typical testing-effect experiment. The point was not to simulate the procedure of a testing experiment; it was to simulate the resulting distributions. The experiment included an initial pool of 80 word pairs. In one condition, the Study-40 condition, forty word pairs were studied once each. In the other condition, twenty of the remaining 40 items were studied twice each and the other 20 items were not studied at all. The Study-40 condition simulates a restudy condition in the sense that all items were all treated equally and their memory strengths were boosted by the roughly the same amount. The Study-20-Twice condition simulates a typical test condition: Twenty of the items were boosted substantially and the other 20 items were not boosted at all. We assumed that studying some items twice and others not at all would bifurcate the distribution in a manner analogous to the way that items being recalled or not recalled on an initial test bifurcates the distribution.

The prediction that follows from the manipulation in Experiment 3 is that more items from the Study-40 condition than from the Study-20-Twice condition should be recalled at a short retention interval, but that the effects of forgetting should appear to be larger for the Study-40 condition, because the items in the Study-20-Twice condition that were recalled initially should be farther above threshold, on average, and thus less likely to become un-recallable. As a result, relatively or absolutely more items should be recalled in the Study-20-Twice condition at a long delay.

In Experiment 3, all 40 items from both conditions were included on the final criterion test (including the 20 pairs that had not been studied), just as they would be in a testing-effect experiment. Half were tested immediately and half were tested after a 2-day delay. The dependent measure was the number of items participants remembered on the final test.

Method

Participants

Twenty-eight participants were recruited using Amazon's Mechanical Turk. They ranged in age from 18 to 60 (mean = 28.4). Twenty participants were female and eight were male. Only participants who lived in the United States were asked to participate. All participants reported speaking English fluently. They were paid $2.00 for participating in the first session of the experiment, which took about 20 min. They received an additional $2.00 upon completing the second session, which took about 5 min.

Materials

The stimuli were 80 word pairs with forward association strengths between .050 and .054, based on Nelson et al.'s (1998) norms (see Kornell & Bjork, 2009, for a list of these materials). Examples include BOWL–PLATE and DECORATE–CAKE.

Design

The experiment employed a 2 × 2 within-participants design. Study Condition had two levels, Study-40 and Study-20-Twice. Retention Interval had two levels: Half of the items were tested immediately and half were tested after 2 days.

Procedure

The experiment was conducted online. Following the instructions, participants completed 80 study trials. Mixed randomly within the 80 trials were 40 pairs that were studied one time each (the Study-40 condition) and 20 other pairs that were studied twice each (the Study-20-Twice condition). On each trial, participants were shown the cue and target word, and below them they were shown the cue word again and asked to copy the target into a box next to the cue. They were given 5 s to do so.

After the study phase, participants were asked to count backward from 547 by 3 s for 15 s. Then they completed the immediate test. Half of the items from each condition were tested. Thus, 20 items from the Study-40 condition were tested and 20 items from the Study-20-Twice condition were tested. In the latter condition, only 10 of the tested items had been presented for study. On each trial, the cue word was presented and participants were asked to type in the target. They were given unlimited time to do so. No feedback was given during the test.

Two days after the first session, participants were contacted via email and asked to participate in the second session. During the second session, all 80 items from session 1 were tested. The analyses were limited to the 40 items that had not been tested on the immediate test. The test trials were conducted in the same way as in session 1.

Results

The bifurcation model predicts that there should be an interaction between study condition and retention interval, because the Study-20-Twice condition creates a bifurcated distribution of 20 strong items and 20 weak (unstudied) items. The data supported this prediction (see Fig. 4). There was a significant effect of Study Condition ($F(1, 27) = 10.93$, $p < .01$, $\eta_p^2 = .29$) and a significant effect of Retention Interval ($F(1, 27) = 149.12$, $p < .0001$, 1, $\eta_p^2 = .85$). Crucially, there was also a significant interaction ($F(1, 27) = 11.64$, $p < .01$, $\eta_p^2 = .30$).

Discussion

Experiment 3 simulated the ways in which tests affect item distributions. The results showed that items studied twice or not at all appeared to be forgotten more slowly than items studied once. The results were consistent with the bifurcation model: Bifurcating the distribution, in the Study-20-twice condition, appeared to prevent forgetting in the absence of any testing. It appears that the test-delay interaction might be better characterized as a differentially-strengthening-a-subset-of-the-items-delay interaction. The results were neither consistent nor inconsistent with the theory that retrieval prevents forgetting, because the study phase of the experiment did not involve any retrieval.

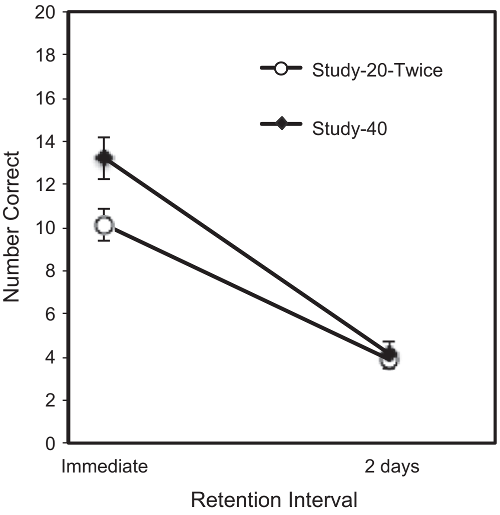

Fig. 4. Number correct in Experiment 3. Eighty items were tested in total, 20 in each condition. Error bars represent ± 1 standard error of the mean.

General discussion

In this article, we have argued that recall tests that are not followed by feedback create bifurcated item distributions in which retrieved items are high in memory strength and unretrieved items are low. Study trials in which all items are treated equally create normal distributions. Items in a bifurcated distribution will tend to cross below the recall threshold less frequently than their untested counterparts, giving the appearance of differential forgetting rates, even if tested and untested items are forgotten at the same rate.

Based on these arguments, which we have called the bifurcation model, we have offered an explanation of the finding, obtained in this article and in a number of others, that recalling information appears to prevent forgetting.

Experiment 1 and 2 showed that compared to a copy condition, retrieval without feedback appears to prevent forgetting. But there were two other retrieval conditions, the test-plus-feedback condition and the consonants condition, in which all items were treated more-or-less equally. Both of these conditions produced rates of forgetting comparable to that of the Copy condition and significantly greater than that of the Test condition (and both were superior to the Test condition overall). Thus, retrieval without bifurcation did not appear to prevent forgetting. In Experiment 3, we simulated the kind of distribution created by testing without feedback, and found the appearance of reduced forgetting, even though the learning phase of the experiment did not involve any retrieval.

These data support the bifurcation model. They do not appear to support the retrieval-prevents-forgetting hypothesis, although this hypothesis remains viable and could be supported by future research. One of its key predictions, however—that retrieval prevents forgetting even in the presence of feedback—has not been supported in previous research, nor was it supported by the experiments presented here. We further discuss the viability of this hypothesis below.

Additional implications

The main goal of the experiments reported here was to test novel predictions of the bifurcation model. Other findings from the current experiments warrant comment, however.

Experiment 2 has two implications that are worth emphasizing. One is the value of feedback. As Fig. 3 shows, recall accuracy in the Test-plus-feedback condition (M = 81%), which produced the best performance both immediately and after a delay, was markedly (and significantly) superior to recall accuracy in the Test condition (M =

53%). The only methodological difference between the two conditions was that, during a mere 2 s after each trial, feedback was provided in the Test-plus-feedback condition but withheld in the Test condition. Thus, if educators and students want to take advantage of the testing effect, they should recognize that incorporating feedback plays an essential role in making tests beneficial (particularly after unsuccessful tests; see Hays, Kornell, & Bjork, 2010; Pashler et al., 2005). A corollary of this conclusion is that in many cases, tests without feedback are not an efficient way to study.

We have tried to emphasize that the bifurcation model and the retrieval-prevents-forgetting model are not mutually exclusive. Both might be at work simultaneously. In Experiment 2, however, if retrieval prevented forgetting, the Test-plus-feedback condition should surely have shown less forgetting than did the Copy condition. It did, but only by about 5 percentage points. (The rate of forgetting, i.e., immediate recall minus delayed recall, was .20 in the Test-plus-feedback condition and .25 in the Copy condition). An ANOVA comparing the Test-plus-feedback and Copy condition showed that the difference in forgetting rates between the two conditions was not significant, $F(1, 39) = 1.38$, $p = .25$. Carpenter et al. (2008) found similar results when they compared forgetting rates for a Test-with-feedback condition and a Study condition: Except when there was a ceiling effect, in their Experiment 2, ANOVAs did not produce significant interactions. (As an alternative to ANOVA, they also did a curve-fitting analysis. This analysis assumed that if some items are more retrievable than others at a given point in time, they are bound to be forgotten at a lower rate, even if the forgetting rates appear to be parallel, because eventually both functions will go to zero. In this analysis, there were differences in forgetting rates across conditions.)

Thus, past research (Carpenter et al., 2008), as well as Experiment 2, both suggest that if retrieval prevents forgetting, the effect is fairly small. Furthermore, any actual difference between the Test-plus-feedback and Copy condition can be explained by the bifurcation model under the assumption that the value of a successful test plus feedback exceeds the value of an unsuccessful test plus feedback. If that assumption is valid, even tests with feedback will bifurcate item distributions, if less dramatically than tests without feedback. The apparent small difference in forgetting rate between the Test-plus-feedback and Copy conditions in Experiment 2, shown in Fig. 3, may reflect such a dynamic.

Experiment 3 also has two implications that are worth noting. When the items in Experiment 3 were tested after a 2-day delay, there was essentially no difference in performance between studying 40 items once and studying 20 items twice. This finding suggests an equivalence between the relative values of restudying an item versus studying a novel item instead. That is, after 20 items had been studied, studying 20 new items was equal in value to restudying the same 20 items again. Apparently, the second of two spaced study trials was just as valuable, at least in terms of apparent learning, as the first (for a review of the spacing effect, see Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006). It should be remembered, however, that the Study-40 condition lost its advantage after 2 days because (presumably) of the bifurcation of the items in the Study-20-Twice condition. Thus, in terms of total memory strength, it is possible that the Study-40 condition was equal to, or even stronger than, the Study-20-Twice condition.

Experiment 3 has a second implication. There is more than one way to bifurcate a distribution. Experiments 1–3 showed that tests do so, as does studying some items more than others. Another way to do so, without using tests, would be to use a mix of easy and difficult items. Such items would naturally create a bifurcated distribution of memory strengths, parallel to retrieved (easy) and unretrieved (difficult) items. A set of items whose difficulty is moderate and normally distributed is parallel to a set of restudied items. A mix of easy and difficult items could appear to prevent forgetting in the same way that tests do.

There is also more than one way to change the position of items relative to the recall threshold. In the present experiments, test difficulty was manipulated by manipulating the retention interval between learning and a final test. Halamish and Bjork (in press) manipulated test difficulty in another way, based on test format (e.g. free recall versus cued-recall). Doing so produced the type of interaction found in the present experiments: Changing the test format was akin to changing the threshold for recall, and as the threshold became more stringent, initial testing became more advantageous, as compared to initial restudying.

The relative importance of memory strength versus recall accuracy

Recall tests differentiate between items above and below a recall threshold. They do not differentiate between two items that are both below threshold, nor between two items that are both above threshold. In the short term, these latter differences do not matter much; the important thing is what can be recalled now. But future learning and forgetting guarantee that the position of an item relative to the recall threshold will change over time. Item strength prevents forgetting and enhances future learning. Thus, in the long term, an item's strength is more important than whether it is above or below threshold at this moment.

The idea that item strength is the key to long-term learning becomes important when making recommendations about how to optimize learning. For example, consider relearning. Relearning can be just as important as initial learning; because it happens so often, it can even be more important. Relearning is easier when an item is just below threshold than when it is very weak. The bifurcation model predicts that when average recall accuracy is higher among tested than restudied items, the restudied items that are below threshold will be stronger than the tested items that are below threshold. As a result, the restudied items will be easier to relearn than tested items. If, in a testing effect experiment, all items were restudied after a delay and then tested, instead of simply being tested after a delay, a relatively large number of previously restudied items could become recallable. Thus, the long-term advantage of test trials could turn into an advantage for presentation trials. In other words, the apparent benefits of tests without feedback might actually turn into costs for long-term relearning (see Storm, Friedman, & Bjork, 2009).

These arguments suggest that choosing appropriate items to study can be complicated. According to the region of proximal learning (RPL) model of study-time allocation, people benefit most from studying the easiest items that they have not already learned (Kornell & Metcalfe, 2006; Metcalfe, 2002; Metcalfe & Kornell, 2003, 2005; Son & Metcalfe, 2000). These low-hanging fruit are easiest to move from below threshold to above threshold. But when one's goal is to master even the difficult items (e.g., if one intends to become fluent in a foreign language), the optimal strategy may be to focus on the most difficult items. Even if doing so is the worst strategy as measured by a relatively short-term recall test, it may be the best way to move all items above threshold permanently.

Concluding comment

Considering item distributions, rather than simply taking averages, is like knowing how to catch a child who is flying through the air: It is usually unnecessary, but there are situations where it can be extremely useful. Bifurcated distributions affect apparent forgetting rates. They also appear to affect metacognitive ratings (Kimball & Metcalfe, 2003; Spellman & Bjork, 1992). Moreover, they might affect practical recommendations about ways to increase learning efficiency. Manipulations that result in the most items being above threshold at a given time produce the greatest recall accuracy at that time. But these manipulations do not necessarily produce the most overall memory strength. Memory strength, not recall accuracy at a particular time, is the key to preventing forgetting and promoting future learning.

Acknowledgments

Elizabeth Bjork, Patricia Jacobs Klein, and Vered Halamish deserve thanks for their significant contributions to this article. We also thank members of Cogfog for intermittent contributions to the distribution framework over the last three decades. Grant 29192G from the McDonnell Foundation supported this research.

Appendix A. Materials used in Experiment 1

Cue	Target	Cue	Target
Antidote	Cure	Grave	Dead
Apple	Fruit	Guide	Lead
Atlas	World	Head	Hair
Average	Normal	Heart	Beat
Barn	Horse	Hour	Time
Bench	Seat	Huge	Large
Birth	Baby	Inhale	Breath
Bridge	River	Just	Fair
Bump	Speed	Kitchen	Cook
Bunch	Group	Learn	Teach
Capital	State	Leave	Come
Champagne	Wine	Letter	Write
Clerk	Store	Lips	Mouth
Clown	Laugh	Living	Room
Coin	Penny	Mask	Hide
Corpse	Body	Mess	Clean
Costume	Party	Motion	Move
Crack	Break	Noose	Rope
Daring	Brave	Observe	Look
Dive	Pool	Orchard	Flower
Domain	Home	Paper	Pencil
Door	Window	Picnic	Basket
Dove	White	Playing	Game
Drip	Water	Polite	Rude
Earth	Planet	Position	Place
Fiction	Book	Prism	Color
First	Second	Puddle	Rain
Flag	Pole	Purpose	Goal
Furniture	Chair	Raft	Boat
Garlic	Bread	Razor	Shave

References

Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In A. F. Healy, S. M. Kosslyn, & R. M. Shiffrin (Eds.). *From learning processes to cognitive processes: Essays in honor of William K. Estes* (Vol. 2, pp. 35–67). Hillsdale, NJ: Erlbaum.

Bjork, R. A., Hofacker, C., & Burns, M. J. (1981). *An "effectiveness-ratio" measure of tests as learning events*. Paper presented at the meetings of the Psychonomic Society, November.

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

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin, 132*, 354–380.

Craik, F. I. M., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General, 104*, 268–294. doi:10.1037/0096-3445.104.3.268.

Cull, W. L. (2000). Untangling the benefits of multiple study opportunities and repeated testing for cued recall. *Applied Cognitive Psychology, 14*, 215–235.

Gelfand, H., Bjork, R. A., & Kovacs, K. E. (1983). *Retrieval as a recognition-memory modifier: A distribution-based theory*. Paper presented at the meetings of the psychonomic society, November.

Halamish, V., & Bjork, R. A. (in press). When does testing enhance retention? A distribution-based interpretation of retrieval as a memory modifier. *Journal of Experimental Psychology: Learning, Memory, & Cognition*.

Hays, M. J., Kornell, N., & Bjork, R. A. (2010). The costs and benefits of providing feedback during learning. *Psychonomic Bulletin & Review, 17*, 797–801. doi:10.3758/PBR.17.6.797.

Hogan, R. M., & Kintsch, K. W. (1971). Differential effects of study and test trials on long-term recognition and recall. *Journal of Verbal Learning and Verbal Behavior, 10*, 562–567.

Izawa, C. (1970). Optimal potentiating effects and forgetting-prevention effects of tests in paired-associate learning. *Journal of Experimental Psychology, 83*, 340–344.

Karpicke, J. D. (2009). Metacognitive control and strategy selection: Deciding to practice retrieval during learning. *Journal of Experimental Psychology: General, 138*, 469–486.

Karpicke, J. D., & Roediger, H. L. (2007). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language, 57*, 151–162.

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

Kimball, D. R., & Metcalfe, J. (2003). Delaying judgments of learning affects memory, not metamemory. *Memory & Cognition, 31*, 918–929.

Kornell, N., & Bjork, R. A. (2009). A stability bias in human memory: Overestimating remembering and underestimating learning. *Journal of Experimental Psychology: General, 138*, 449–468.

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

Kornell, N., & Metcalfe, J. (2006). Study efficacy and the region of proximal learning framework. *Journal of Experimental Psychology: Learning Memory, & Cognition, 32*, 609–622.

Metcalfe, J. (2002). Is study time allocated selectively to a region of proximal learning. *Journal of Experimental Psychology: General, 131*, 349–363.

Metcalfe, J., & Kornell, N. (2003). The dynamics of learning and allocation of study time to a region of proximal learning. *Journal of Experimental Psychology: General, 132*, 530–542.

Metcalfe, J., & Kornell, N. (2005). A region of proximal learning model of study time allocation. *Journal of Memory and Language, 52*, 463–477.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). The University of South Florida word association, rhyme, and word fragment norms. <http://w3.usf.edu/FreeAssociation/>.

Pashler, H., Cepeda, N. J., Wixted, J. T., & Rohrer, D. (2005). When does feedback facilitate learning of words? *Journal of Experimental Psychology: Learning, Memory, & Cognition, 31*, 3–8.

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

Richland, L. E., Kornell, N., & Kao, L. S. (2009). The pretesting effect: Do unsuccessful retrieval attempts enhance learning? *Journal of Experimental Psychology: Applied, 15*, 243–257.

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.

Runquist, W. (1986). The effect of testing on the forgetting of related and unrelated associates. *Canadian Journal of Psychology, 40*, 65–76.

Slamecka, N. J., & Katsaiti, L. T. (1988). Normal forgetting of verbal lists as a function of prior testing. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 14*, 716–727.

Son, L. K., & Metcalfe, J. (2000). Metacognitive and control strategies in study-time allocation. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 26*, 204–221.

Spellman, B. A., & Bjork, R. A. (1992). When predictions create reality: Judgments of learning may alter what they are intended to assess. *Psychological Science, 3*, 315–316.

Storm, B. C., Friedman, M. C., & Bjork, R. A. (2009). On the transfer of prior tests or study events to subsequent study. *Poster presented at the 50th annual meeting of the psychonomic society, November*, Boston, Massachusetts.

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.

Toppino, T. C., & Cohen, M. S. (2009). The testing effect and the retention interval: Questions and answers. *Experimental Psychology, 56*, 252–257. doi:10.1027/1618-3169.56.4.252.

Tulving, E. (1967). The effects of presentation and recall of material in free-recall learning. *Journal of Verbal Learning and Verbal Behavior, 6*, 175–184.

Wheeler, M. A., Ewers, M., & Buonanno, J. E. (2003). Different rates of forgetting following study versus test trials. *Memory, 11*, 571–580.