Journal of Experimental Psychology:
Learning, Memory, and Cognition
2013, Vol. 39, No. 1, 290–296

0278-7393/12/$12.00 DOI: 10.1037/a0028468

RESEARCH REPORT

When and Why a Failed Test Potentiates the Effectiveness of Subsequent Study

Matthew Jensen Hays
University of Southern California

Nate Kornell
Williams College

Robert A. Bjork
University of California, Los Angeles

Teachers and trainers often try to prevent learners from making errors, but recent findings (e.g., Kornell, Hays, & Bjork, 2009) have demonstrated that tests can potentiate subsequent learning even when the correct answer is difficult or impossible to generate (e.g., "What is Nate Kornell's middle name?"). In 3 experiments, we explored when and why a failed test enhances learning. We found that failed tests followed by immediate feedback produced greater retention than did a presentation-only condition. Failed tests followed by delayed feedback, by contrast, did not produce such a benefit—except when the direction of the final test was reversed (i.e., the participants were provided with the target and had to produce the original cue). Our findings suggest that generating an incorrect response to a cue both activates the semantic network associated with the cue and suppresses the correct response. These processes appear to have 2 consequences: If feedback is presented immediately, the semantic activation enhances the mapping of the cue to the correct response; if feedback is presented at a delay, the prior suppression boosts the learning of the suppressed response.

Keywords: spacing, testing, retrieval, forgetting, errorless learning

Tests are perhaps the most ubiquitous element of formal education. For hundreds of years, almost every lecture, lab, and seminar has concluded with a test. These criterion tests are intended to diagnose the knowledge or ability possessed by the student or trainee. In the past century, however, researchers have discovered that tests do more than diagnose; they are potent learning events (e.g., Gates, 1917; Spitzer, 1939). That is, tests serve to create and/or strengthen associations between information in memory. Indeed, successfully retrieving information during learning often creates stronger memories than does re-studying (e.g., Allen, Mahler, & Estes, 1969). This testing effect has been demonstrated in a variety of domains, including children's picture naming (Wheeler & Roediger, 1992), high-school students' comprehension of history lessons (Nungester & Duchastel, 1982), college students' comprehension of cognitive psychology course material (Leeming, 2002), and undergraduate students' recollection of idea units from prose passages (Roediger & Karpicke, 2006).

Nevertheless, tests are rarely used to improve learners' comprehension or competence in mainstream education (e.g., Glover, 1989). The threat of a test is often used to encourage students to study more (but with little effect on learning; see Haynie, 1997). Educators may be reluctant to employ tests as learning events for fear of the effects of incorrect responses. This fear is not without substance. Butler and Peterson (1965) found that errors on tests can be "stamped in," meaning that producing an incorrect response will cause that same incorrect response to be produced on later tests (but see Metcalfe & Kornell, 2007). These findings reinforced the emerging "errorless learning" movement (e.g., Terrace, 1963), which posited that errors weakened instruction and created unwanted by-products of the training process. Indeed, errorless learning does appear to benefit clinical populations (e.g., Kern, Liberman, Kopelowicz, Mintz, & Green, 2002; Kessels & de Haan, 2003).

In normal classrooms and training settings, however, research suggests that tests should still be used—even if students sometimes respond incorrectly (e.g., Marsh, Roediger, Bjork, & Bjork, 2007; Pashler, Zarow, & Triplett, 2003). Failed tests, it seems, do not overwrite previously learned information or otherwise corrupt cognition. Further, failed tests do not reduce the value of later learning. On the contrary, Kornell et al. (2009) found that presen-

This article was published Online First May 14, 2012.

Matthew Jensen Hays, Institute for Creative Technologies, University of Southern California; Nate Kornell, Department of Psychology, Williams College; Robert A. Bjork, Department of Psychology, University of California, Los Angeles.

Grant 29192G from the McDonnell Foundation supported this research. We thank the members of CogFog for their suggestions and interpretations of the findings reported in this article. We especially thank Barbara Knowlton and John Nestojko for critical methodological suggestions throughout our research.

Correspondence concerning this article should be addressed to Matthew Jensen Hays, USC–ICT, 12015 Waterfront Drive, Los Angeles, CA 90094. E-mail: matt@hayslab.com

tations of information were more powerful when they were preceded by failed attempts to produce that information.

Kornell et al. (2009) avoided item-selection effects by using cues that made success impossible or highly unlikely. In Experiments 1 and 2, they used trivia questions (e.g., "What peace treaty ended the Calumet War?") invented by Berger, Hall, and Bahrick (1999). Their participants believed that these questions were real (albeit obscure and difficult) especially because they were mixed in with genuine trivia questions. Their participants therefore attempted to answer the invented questions, but they had no chance to produce the "correct" answer (e.g., "Harris Treaty"). Their responses were deemed "incorrect," and they were subsequently given immediate corrective feedback: namely, the intact invented question–answer pair. Their participants recalled more such items on the final test than did participants who were allotted the same amount of time to study the intact pair without having generated a wrong answer (Experiment 1), and they recalled as many items as did participants who, without having generated a wrong answer, were given more than twice as much time to study the intact question–answer pair (Experiment 2).

In Experiments 3–6, Kornell et al. (2009) used pairs of words with weak semantic associations (called *low associates*). Before viewing the intact pairs (e.g., "frog–pond"), their participants sometimes saw the first word (the *cue*) and attempted to produce the second word (the *target*). The participants produced a different associate of the cue, such as "toad," about 97% of the time and then received immediate corrective feedback. These items were recalled reliably more often on the final test than were items that did not receive a test but were instead displayed intact for longer. This advantage of failed tests was manifested whether the final test was administered at the end of the learning phase or after a 24-hr delay.

Richland, Kornell, and Kao (2009) also reported that failed tests enhance memory. They tested their participants on some information in a two-page neuroscience text that they had not yet read. The participants, of course, were unable to produce the correct responses. After they read the text, they received a final test that included the earlier questions. Richland et al. found that the facts on which the participants had been tested were recalled at a substantially higher rate than were corresponding facts on which they had not been tested.

Further evidence for the value of failed tests comes from studies in which the number of tests has been manipulated. Multiple failures to produce the correct response led to better recall than did a single failure (Izawa, 1966, 1967). Related studies (e.g., Benjamin, Bjork, & Schwartz, 1998; Gardiner, Craik, & Bleasdale, 1973; Whitten & Bjork, 1977) have shown that participants' final-test recall was better for items that they had to work longer to produce during learning (i.e., on which they had to expend more effort). Auble and Franks (1978) directly manipulated the duration of this attempt and also found effort to be positively correlated with the probability of final-test recall. Together, these findings indicate the potentiating effect of tests; tests intensify the learning that occurs when the answer is finally made available (e.g., Izawa, 1970, 1971).

Kornell et al. (2009) suggested that the potentiating effects of a (failed) test might reflect the semantic activation triggered by the test. Thus, for example, the effort expended in anticipating the target of a particular cue (e.g., "frog–___") activates the semantic network of information related to that cue. When the answer becomes available (either via retrieval or via external feedback), this priming allows it to be efficiently mapped to the cue (e.g., frog–pond). This mapping strengthens the cue–target association, thereby slowing forgetting and increasing the probability that the pair can be completed on the final test.

The Present Study

The present study was intended to test the idea that persisting semantic activation from a failed test can enhance subsequent learning—and that this enhancement fades over time. Our approach was to manipulate the amount of time that separated the test and the feedback event. Specifically, we measured recall after varying whether a failed test was followed by immediate feedback (i.e., massed) or delayed feedback (i.e., spaced).

Typically, separating two presentations of information is beneficial for long-term memory. This spacing effect has been demonstrated hundreds of times. Cepeda, Pashler, Vul, Wixted, and Rohrer (2006) reviewed dozens of reports of spacing effects in the verbal learning literature alone. In the case of failed tests, although the incorrect initial response differs from the subsequent feedback, perhaps delaying feedback would improve recall. Indeed, delayed feedback often results in greater retention than does immediate feedback (e.g., Smith & Kimball, 2010).

On the contrary, according to Kornell et al.'s (2009) hypothesis, the to-be-learned answer must become available before the activation triggered by the test dissipates, or the test will not enhance subsequent encoding. Thus, the beneficial effects of a failed test should decrease when spacing separates a failed test from the answer. We therefore predicted an advantage for immediate feedback over delayed feedback.

Like Kornell et al. (2009, Experiments 3–6), we used low associates (e.g., "frog–pond"). Experiment 1 compared recall for word pairs in a 2 × 2 design. Pairs were either tested and then presented (test–present) or presented and then presented again (present–present). The two trials on a given item were either consecutive (massed) or separated by other items (spaced). Experiment 2 compared recall in the two test–present conditions—massed test–present and spaced test–present—against a single presentation of the word pair (also see Grimaldi & Karpicke, in press). Experiment 3 reversed the direction of the final test to more thoroughly examine the effects of the failed test on the cue–target association. (We also report the results of an additional experiment in which we asked the participants to free-recall the cues. The result of this experiment allowed us to eliminate cue memorability as an explanation for the pattern of results in Experiments 1–3.)

Experiment 1

Method

Participants. The participants were 70 undergraduates at the University of California, Los Angeles (UCLA). They received partial course credit as compensation for their participation.

Materials, procedure, and design. The materials were 60 word pairs with weak semantic associations (e.g., "frog–pond"). Each word was a minimum of four letters long. The forward association strength of each pair was between .050 and .054; the

average was .052 (Nelson, McEvoy, & Schreiber, 1998). In other words, when shown the cue, people produce the correct target as their first response approximately 5.2% of the time.

Half of the pairs were randomly assigned to receive a test (e.g., "frog–___") followed by a presentation (e.g., "frog–pond"). The other half of the items were presented twice. In both cases, the first event (test or presentation) was 8 s long and the second event (presentation) was 5 s long.

These two types of items were then equally and randomly distributed among three schedules: massed, spaced, or filler. These schedules determined when in the learning period the participants would encounter the first event and the second event. On a spaced schedule, the two events were separated by approximately 9.5 min. This separation was achieved by scheduling the first event in the first half of the learning period and the second event in the second half of the learning period. On a massed schedule, the first and second events were encountered consecutively in the second half of the learning period. Filler items' first and second events were encountered consecutively in the first half of the learning period.

Thus, the first half of the learning period comprised all filler items as well as the first event for spaced items. The second half of the learning period comprised all massed items as well as the second event for spaced items. The average interval between the second event and the final test was therefore equated in the massed and spaced conditions. Filler items were omitted from all analyses. The four conditions of interest, thus, were massed test–present, spaced test–present, massed present–present, and spaced present–present; the experiment used a 2 (item type) × 2 (schedule) within-subjects factorial design.

After a 5-min-long distractor (an unrelated visuospatial task), all items were tested (e.g., "frog–___") in random order. The dependent variable was final-test recall. Following Kornell et al. (2009), we omitted from all analyses the very rare item(s) for which a participant provided the correct response on the initial test.

Results and Discussion

Figure 1 presents the participants' ability to recall the target when they were provided with the cue on the final test. There was no main effect of schedule on recall. The participants recalled approximately as many spaced items ($M = .68$, $SE = .02$) as massed items ($M = .69$, $SE = .02$), $F < 1$, *ns*. There was a main effect of item type on recall. The participants' recall was greater for present–present items ($M = .71$, $SE = .02$) than for test–present items ($M = .66$, $SE = .02$), $F(1, 69) = 7.46$, $p < .01$, $\eta^2 = .10$.

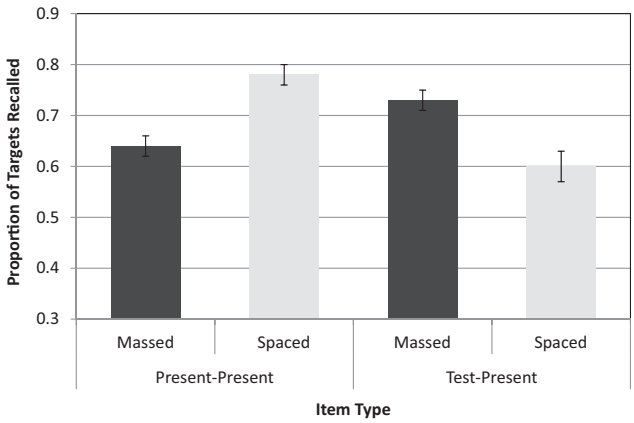

Figure 1. Mean final-test recall by condition in Experiment 1. Error bars represent the standard error of the mean.

These main effect findings were qualified by a statistically significant interaction between item type and schedule on recall, $F(1, 69) = 49.91$, $p < .001$, $\eta^2 = .42$. As can be seen in Figure 1, recall was greater for massed test–present items than for massed present–present items, $t(69) = 3.55$, $p < .01$. Immediately preceding a presentation with a failed test led to better recall than did immediately preceding it with another presentation. This finding replicates the results of Kornell et al. (2009).

A delay, however, reversed this pattern. Recall was greater for spaced present–present items than for spaced test–present items, $t(69) = 7.00$, $p < .001$. At a delay, preceding a presentation with a failed test led to *worse* recall than did preceding it with another presentation. This finding suggests that the effects of a failed test are very different from the effects of a presentation or a successful test. Whereas a delay before feedback enhances learning following a successful test (e.g., Cepeda et al., 2006), a failed test appears to be even less useful than a presentation. This reversal of the testing effect is consistent with our notion that the beneficial effect of a test dissipates over time.

Figure 1 also shows that recall was greater for spaced present–present items than for massed present–present items, $t(69) = 7.38$, $p < .001$. This advantage of spacing over massing is consistent with a considerable body of literature on the spacing effect. Recall was greater, however, for massed test–present items than for spaced test–present items, $t(69) = 4.51$, $p < .001$.

The above results were not likely due to a selection effect caused by differential success rates on the initial test. During learning, the participants provided the correct response on approximately as many massed test–present items ($M = .04$, $SE = .01$) as spaced test–present items ($M = .04$, $SE = .01$), $t(69) < 1$, *ns*.

The overall pattern of results in Experiment 1 is remarkable. The testing effect and spacing effect are two of the most consistent and powerful phenomena in the cognitive psychology literature. Nevertheless, neither spacing nor testing appears to be beneficial when the initial learning event is a failed test. Instead, a failed test seems to potentiate later learning, but this potentiation fades as time separates the failed test from the subsequent presentation. The rate at which this potentiation fades is impossible to determine from the results of Experiment 1. Indeed, there is no way to determine from the results of Experiment 1 whether delayed feedback after failed tests has any benefit whatsoever. Experiment 2 was designed to address this issue.

Experiment 2

Determining how failed tests affect subsequent presentations is critical to the present goal of understanding the consequences of tests themselves on learning. Specifically, we wanted to determine whether a failed test has any effect on the learning that occurs during a delayed subsequent presentation. Thus, in Experiment 2, we included a single-presentation condition. This condition allowed us to compare items that were presented alone versus items that were presented as delayed feedback following an unsuccessful test. Thus, in Experiment 2, recall for massed test–present and

spaced test–present items was compared with recall for word pairs that were presented only once.

Method

Participants. The participants were 45 UCLA undergraduates. They received partial course credit as compensation for their participation.

Materials, procedure, and design. The materials used in Experiment 2 were identical to those used in Experiment 1. The present–present conditions used in Experiment 1 (both massed and spaced) were removed. A single-presentation condition was added in their place: The intact pair was displayed for 5 s. (Five seconds was also the duration of the display of feedback in the other two conditions.) Thus, the three conditions of interest were massed test–present, spaced test–present, and single-presentation.

The first half of the learning period included the test for spaced test–present items. It also included filler items that corresponded to the other two conditions of interest. The second half of the learning period included the presentation for spaced test–present items, the test and the presentation for massed test–present items, and all single-presentation items.

Results and Discussion

Figure 2 presents the participants' ability to recall the target when provided with the cue on the final test. There was a main effect of item type, $F(2, 88) = 8.74$, $p < .001$, $\eta^2 = .17$. The participants recalled more massed test–present items than single-presentation items, $t(44) = 4.37$, $p < .001$. This finding is consistent with prior work in which adding a failed test before a presentation enhanced learning (Kornell et al., 2009; Richland et al., 2009). Figure 2 also shows that the participants recalled more massed test–present items than spaced test–present items, $t(44) = 3.12$, $p < .01$. This finding is consistent with the results of Experiment 1. Also as in Experiment 1, this finding is not likely due to a selection effect caused by differential success rates on the initial test. During learning, the participants provided the correct response on approximately as many massed test–present items ($M = .05$, $SE = .01$) as spaced test–present items ($M = .04$, $SE = .01$), $t(44) < 1$, *ns*.

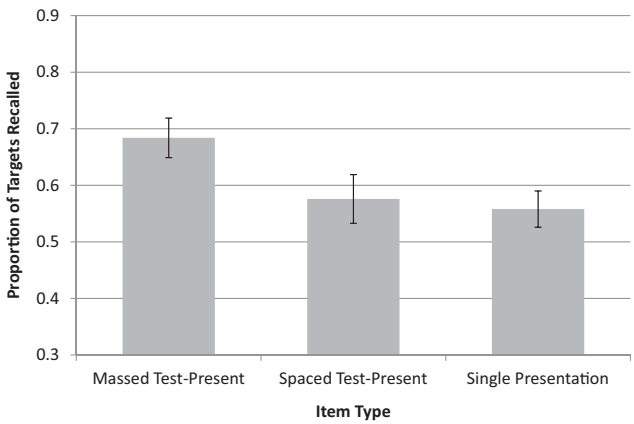

Figure 2. Mean final-test recall by condition in Experiment 2. Error bars represent the standard error of the mean.

The key question in Experiment 2 was whether participants would recall more spaced test–present items than single-presentation items. They did not ($t < 1$, *ns*). At a delay, failed tests appear to have little measurable effect. That is, they were not harmful, but they were also not helpful. This result provides further support for the notion that the beneficial effects of a failed test wear off over time.

However, these results appear to be at odds with the findings of Richland et al. (2009). In their experiments, the delay between failed tests and presentations was also a few minutes—but they found powerful and persistent beneficial effects of the failed tests. Perhaps there were beneficial effects of the failed test in the present study—but we were unable to detect them. A possible explanation is that the beneficial effects were obscured by one or more sources of interference.

Evidence of this interference can be inferred from the sources and frequency of errors on the initial and final tests. On the initial test, the participants (incorrectly) provided the strongest associate of the cue on approximately 17% of trials ($SE = .02$). On the final test, they provided their initial incorrect response far more often ($M = .40$, $SE = .05$) than they provided the strongest associate ($M = .09$, $SE = .01$). This disparity suggests that the initial test created a powerful lure for the participants on the final test. Thus, the negative consequences of the initial test may have rendered the positive consequences unable to be detected. Experiment 3 was designed to avoid the negative consequences of the initial test in order to determine whether there had been undetected positive consequences.

Experiment 3

Experiment 2 suggested that tests may have no value if they are followed by delayed feedback. It is possible, however, that interference may have obscured the beneficial effects of a failed test on delayed feedback. Experiment 3 was designed to circumvent this interference. Again, our goal was to compare the effects of an initial failed test on a (immediate or delayed) presentation versus a single presentation with no prior test. To that end, Experiment 3 was identical to Experiment 2, except for one key change: On the final test in Experiment 3, the participants were provided with the target and attempted to produce the cue (e.g., "___–pond"). In this way, competition from other associations to "frog" (e.g., toad) should be reduced or eliminated.

Method

Participants. The participants were 52 UCLA undergraduates. They received partial course credit as compensation for their participation.

Materials, design, and procedure. The materials, design, and procedure used in Experiment 3 were identical to those used in Experiment 2, with one exception. Whereas the final test in Experiment 2 prompted the participants with the cue (e.g., "pond–___"), the final test in Experiment 3 prompted the participants with the target (e.g., "___–frog"). (As in Experiments 1 and 2, trials during the learning period displayed the cue and required the participants to produce the target.)

The average backward association strength of each pair was .046 (Nelson et al., 1998). In other words, when shown the target, people produce the cue approximately 4.6% of the time.

Results and Discussion

Figure 3 presents the participants' ability to produce the cue when they were provided with the target on the final test. There was a main effect of item type, $F(2, 102) = 43.02$, $p < .001$, $\eta^2 = .46$. The participants recalled more massed test–present items than spaced test–present items, $t(51) = 2.80$, $p < .01$. Thus, immediate feedback again improved recall compared with delayed feedback. This finding is consistent with the results of Experiments 1 and 2; the spacing effect is reversed when the first event is a failed test. As in Experiments 1 and 2, this finding is not likely due to a selection effect caused by differential success rates on the initial test. During learning, the participants provided the correct response on approximately as many massed test–present items ($M = .04$, $SE = .01$) as spaced test–present items ($M = .04$, $SE = .01$), $t(51) < 1$, *ns*.

The participants also recalled more massed test–present items than single–presentation items, $t(51) = 10.42$, $p < .001$. This finding is consistent with the results of Experiment 2 and provides further support for the findings of Kornell et al. (2009). This result also shows that the strengthening provided by the failed test is not specific to the test format. That is, even though the participants attempted to provide the target during the learning period, this attempt improved their later ability to provide the cue on the final test.

Critically, the participants recalled more spaced test–present items than single-presentation items, $t(51) = 5.88$, $p < .001$. One may suppose that this advantage is an artifact of the test format. The participants were tested on the cue—which had been seen twice in the test conditions but only once in the single-presentation condition. This additional exposure, and the resultant additional cue memorability, could have been responsible for the relative advantage of the test conditions over the single-presentation condition.

We performed an additional experiment in which we again changed the final-test format—this time, to be a free-recall test of all cues. If the test-format artifact was responsible for our findings in Experiment 3, we would have again found an advantage for massed test–present over spaced test–present. Instead, we found the opposite pattern of results: a standard spacing effect. Thus, cue memorability cannot explain the pattern of results in Experiment 3; there appear to be effects of failed tests that specifically affect the association between cue and target. (Indeed, given the advantage of the spaced test–present condition, cue memorability may play a role in obscuring the effect, rather than contributing to it.)

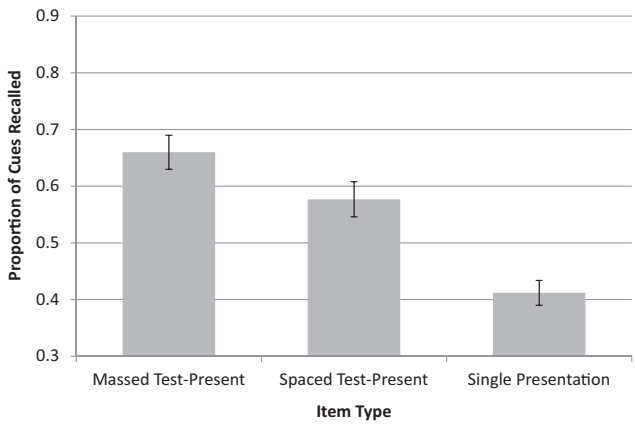

Figure 3. Mean final-test recall by condition in Experiment 3. Error bars represent the standard error of the mean.

It therefore appears that the initial test had a beneficial effect, even after a delay, but that it could not be detected in Experiments 1 and 2 because of interference. By reversing the direction of the test in Experiment 3, we circumvented this interference. As we explore in the next section, though, the results of all three experiments together pose a theoretical puzzle of sorts. They suggest that the semantic-activation explanation of the superiority of the massed test–present condition over the spaced test–present condition may not provide a complete account of the processes engaged when a failed test is followed by feedback.

General Discussion

In three experiments, we demonstrated that failed tests potentiate learning from later presentations. This demonstration served to replicate previous findings (Kornell et al., 2009; Richland et al., 2009). Experiment 3 further revealed that the cue–target association was being potentiated; the benefit was present even when—unlike the failed test—the final test provided the target and required the participants to produce the cue. We also found that the benefit of an unsuccessful test was diminished when a delay separated it from its feedback.

The results of the first two experiments are consistent with the notion that a test serves to prime knowledge related to the cue. In the case of semantic word pairs (e.g., the present study; Kornell et al., 2009, Experiments 3–6), associations between the cue and other semantically related words might be primed. For example, "frog–___" may activate *frog–pond*, *frog–amphibian*, and *frog–tadpole*. In the case of trivia questions (e.g., Kornell et al., Experiments 1 and 2) or passages (Richland et al., 2009), a broad network of semantically relevant information might be primed. For example, "What peace treaty ended the Calumet War?" might activate knowledge of battles, treaties, protestors, war movies, theater, other conflicts, related historical events, and more. Then, when feedback is provided, this activation facilitates the mapping of the cue to the target. Critically, this activation wears off over time. As a result, failed tests are more beneficial when they are followed by immediate feedback than when they are followed by delayed feedback. This explanation is similar to the conclusion reached by Grimaldi and Karpicke (in press), who also investigated the effect of failed tests on immediate and delayed feedback.

However, the pattern of results in Experiment 3 is not entirely consistent with this explanation. On the final test in Experiment 3, items in the spaced test–present condition were recalled reliably more frequently than were items in the single-presentation condition. Thus, an initial test potentiated presentations that occurred after several minutes, which were filled with dozens of intervening items. Priming, though, tends to be very short-lived—on the order of a few seconds or intervening items (McNamara, 2005). It is therefore unlikely that priming was responsible for the advantage

of the delayed test–present condition over the single-presentation condition in Experiment 3. Also, if the spaced test–present condition does create proactive interference, and if the benefits of semantic activation are long gone by the time of the presentation of the correct answer in that condition, why was recall in that condition not worse than recall in the single-presentation condition in Experiment 2? It follows that priming may not be the sole mechanism underlying the benefit of failed tests on subsequent presentations.

An alternative explanation, one that was alluded to by Kornell et al. (2009), is that—during the failed test—the participants searched their semantic networks, thereby activating one or more incorrect semantic associates of the cue. By strengthening these cue–competitor associations (e.g., Slamecka & Graf, 1978), the participants may have temporarily suppressed the cue–target association (Anderson, Bjork, & Bjork, 1994). This suppression has been demonstrated to last at least 20 min (Anderson et al., 1994). Crucially, when an association has been suppressed, subsequent relearning events involving that association can become more potent (Bjork & Bjork, 1992; Storm, Bjork, & Bjork, 2008). Thus, in this view, corrective feedback would have provided a sizable benefit to the previously suppressed cue–target association. For example, "frog–toad" being activated and "frog–pond" being suppressed would, paradoxically, enhance the learning of "frog–pond" when it was presented as feedback. This enhanced learning would occur in both the immediate- and delayed-feedback conditions.

Our speculation, then, is that the spaced test–present condition in Experiment 2 ended up producing a level of final cued recall roughly equivalent to the single-presentation condition because a negative factor (proactive interference from incorrect responses generated on the test) was offset by a positive factor (enhanced learning of the correct response, which was suppressed during the earlier generation of an incorrect response). Thus, although the results of the present study and the spacing effect appear to diverge, they may share a common root: Forgetting enhances later learning (Bjork & Bjork, 1992). In the case of failed tests, the cue–target association is weakened via suppression, whereas in the case of spacing, the association is weakened by various sources of interference (e.g., from intervening items). In both cases, the initial impairment yields later enrichment. From this perspective, our results are consistent with current theory and provide a mechanistic explanation for the claim that tests in general potentiate subsequent study (e.g., Izawa, 1970).

Concluding Comment

The present study provides strong evidence that tests have value even when learners do not provide the correct responses—indeed, even when they cannot possibly succeed—as long as feedback is supplied. Incorrect responses do not indelibly stain students' memories. Instead, with well-timed feedback, tests can actually provide the opportunity for more powerful learning.

References

Allen, G. A., Mahler, W. A., & Estes, W. K. (1969). Effects of recall tests on long-term retention of paired associates. *Journal of Verbal Learning and Verbal Behavior, 8,* 463–470. doi:10.1016/S0022-5371(69)80090-3

Anderson, M. C., Bjork, R. A., & Bjork, E. L. (1994). Remembering can cause forgetting: Retrieval dynamics in long-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 1063–1087. doi:10.1037/0278-7393.20.5.1063

Auble, P. M., & Franks, J. J. (1978). The effects of effort toward comprehension on recall. *Memory & Cognition, 6,* 20–25. doi:10.3758/BF03197424

Benjamin, A. S., Bjork, R. A., & Schwartz, B. L. (1998). The mismeasures of memory: When retrieval fluency is misleading as a metamnemonic index. *Journal of Experimental Psychology: General, 127,* 55–68. doi:10.1037/0096-3445.127.1.55

Berger, S. A., Hall, L. K., & Bahrick, H. P. (1999). Stabilizing access to marginal and submarginal knowledge. *Journal of Experimental Psychology: Applied, 5,* 438–447. doi:10.1037/1076-898X.5.4.438

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

Butler, D. C., & Peterson, D. E. (1965). Learning during "extinction" with paired associates. *Journal of Verbal Learning and Verbal Behavior, 4,* 103–106. doi:10.1016/S0022-5371(65)80092-5

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. doi:10.1037/0033-2909.132.3.354

Gardiner, J. M., Craik, F. I. M., & Bleasdale, F. A. (1973). Retrieval difficulty and subsequent recall. *Memory & Cognition, 1,* 213–216. doi:10.3758/BF03198098

Gates, A. I. (1917). Recitation as a factor in memorizing. In R. S. Woodworth (Ed.), *Archives of psychology* (Vol. 40, pp. 1–104). New York, NY: The Science Press.

Glover, J. A. (1989). The "testing" phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology, 81,* 392–399. doi:10.1037/0022-0663.81.3.392

Grimaldi, P. J., & Karpicke, J. D. (in press). When and why do retrieval attempts enhance subsequent encoding? *Memory & Cognition.*

Haynie, W. J. (1997). Effect of anticipation of tests on delayed retention learning. *Journal of Technology Education, 9,* 20–30.

Izawa, C. (1966). Reinforcement-test sequences in paired-associate learning. *Psychological Reports, 18,* 879–919. doi:10.2466/pr0.1966.18.3.879

Izawa, C. (1967). Function of test trials in paired-associate learning. *Journal of Experimental Psychology, 75,* 194–209. doi:10.1037/h0024971

Izawa, C. (1970). Optimal potentiating effects and forgetting-prevention effects of tests in paired-associate learning. *Journal of Experimental Psychology, 83,* 340–344. doi:10.1037/h0028541

Izawa, C. (1971). The test trial potentiating model. *Journal of Mathematical Psychology, 8,* 200–224. doi:10.1016/0022-2496(71)90012-5

Kern, R. S., Liberman, R. P., Kopelowicz, A., Mintz, J., & Green, M. F. (2002). Applications of errorless learning for improving work performance in persons with schizophrenia. *American Journal of Psychiatry, 159,* 1921–1926. doi:10.1176/appi.ajp.159.11.1921

Kessels, R. P. C., & de Haan, E. H. F. (2003). Implicit learning in memory rehabilitation: A meta-analysis on errorless learning and vanishing cues methods. *Journal of Clinical and Experimental Neuropsychology, 25,* 805–814. doi:10.1076/jcen.25.6.805.16474

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. doi:10.1037/a0015729

Leeming, F. C. (2002). The exam-a-day procedure improves performance in psychology classes. *Teaching of Psychology, 29,* 210–212. doi:10.1207/S15328023TOP2903_06

Marsh, E. J., Roediger, H. L., Bjork, R. A., & Bjork, E. L. (2007).

Memorial consequences of multiple-choice testing. *Psychonomic Bulletin & Review, 14,* 194–199. doi:10.3758/BF03194051

McNamara, T. P. (2005). *Semantic priming: Perspectives from memory and word recognition.* New York, NY: Psychology Press. doi:10.4324/9780203338001

Metcalfe, J., & Kornell, N. (2007). Principles of cognitive science in education: The effects of generation, errors, and feedback. *Psychonomic Bulletin & Review, 14,* 225–229. doi:10.3758/BF03194056

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

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

Pashler, H., Zarow, G., & Triplett, B. (2003). Is temporal spacing of tests helpful even when it inflates error rates? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29,* 1051–1057. doi:10.1037/0278-7393.29.6.1051

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. doi:10.1037/a0016496

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17,* 249–255. doi:10.1111/j.1467-92802006.01693.x

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory, 4,* 592–604. doi:10.1037/0278-7393.4.6.592

Smith, T. A., & Kimball, D. R. (2010). Learning from feedback: Spacing and the delay-retention effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36,* 80–95. doi:10.1037/a0017407

Spitzer, H. F. (1939). Studies in retention. *Journal of Educational Psychology, 30,* 641–656. doi:10.1037/h0063404

Storm, B. C., Bjork, E. L., & Bjork, R. A. (2008). Accelerated relearning after retrieval-induced forgetting: The benefit of being forgotten. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34,* 230–236. doi:10.1037/0278-7393.34.1.230

Terrace, H. S. (1963). Discrimination learning with and without "errors." *Journal of the Experimental Analysis of Behavior, 6,* 1–27. doi:10.1901/jeab.1963.6-1

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. doi:10.1111/j.1467-9280.1992.tb00036.x

Whitten, W. B., & Bjork, R. A. (1977). Learning from tests: Effects of spacing. *Journal of Verbal Learning and Verbal Behavior, 16,* 465–478. doi:10.1016/S0022-5371(77)80040-6

Received October 15, 2011
Revision received February 24, 2012
Accepted March 21, 2012 ■