Title
When does feedback facilitate learning of words?

Journal
Journal of Experimental Psychology-Learning Memory and Cognition, 31(1)

ISSN
0278-7393

Authors
Pashler, Harold
Cepeda, N J
Wixted, J T
et al.

Publication Date
2005

Peer reviewed

Journal of Experimental Psychology: Learning, Memory, and Cognition 2005, Vol. 31, No. 1, 3–8

When Does Feedback Facilitate Learning of Words?

Harold Pashler, Nicholas J. Cepeda, and John T. Wixted
University of California, San Diego

Doug Rohrer
University of South Florida

The question of what form of feedback best promotes associative learning and retention is of obvious practical import. However, the literature on feedback is confusing, with some researchers suggesting that although feedback may enhance performance during learning, it does so at the expense of later retention. To shed further light on this issue, subjects (N = 258) learned Luganda–English word pairs. After 2 initial exposures to the materials, subjects were tested on each item several times, with the presence and type of feedback varying between subjects. Subjects were given a final test on the same items 1 week later. Supplying the correct answer after an incorrect response not only improved performance during the initial learning session—it also increased final retention by 494%. On the other hand, feedback after correct responses made little difference either immediately or at a delay, regardless of whether the subject was confident in the response. Practical and theoretical implications are discussed.

Despite more than a century of work, research on learning and memory has provided designers of classroom curricula or computer-aided instruction systems with surprisingly few bits of concrete guidance on how to speed learning and retard forgetting. This is true even for rather cut and dry learning situations in which people merely seek to acquire discrete bits of information such as facts, foreign language vocabulary, and the like. In part, this lack of translation from basic research to practical application may reflect the fact that, especially in recent years, concrete procedural variables such as temporal distribution of study time, type of testing, and type of feedback have been little studied.

In the present article, we examine one particularly concrete procedural variable, namely, feedback. We ask a seemingly simple question: When a learner has attempted to retrieve discrete information in some sort of cued recall situation (*drill*), what kind of feedback should be provided to maximize what the learner will be able to remember after a delay? The effect of feedback was studied in the 1960s and 1970s and has been discussed in some influential recent reviews, but (we argue) this basic empirical question remains quite unresolved. Below, we describe an experiment in which we look at foreign language vocabulary learning and compare several different forms of feedback, assessing their impact on both immediate learning and a delayed test of retention.

Correspondence concerning this article should be addressed to Harold Pashler, Department of Psychology, 0109, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0109. E-mail: hpashler@ucsd.edu

Research and Theory on Feedback

For most people, common sense would suggest that providing feedback is bound to be useful. After all, it may allow incorrect mental contents to be repaired or replaced, and useful mental linkages to be strengthened. It is surprising, however, that a number of recent reviews have argued that although feedback (and more specifically, advising the learner about exactly what response he or she should have made on a previous trial) may well improve performance during training, it often does so at the expense of longer term retention (e.g., Bjork, 1994; Rosenbaum, Carlson, & Gilmore, 2000; Schmidt & Bjork, 1992). Similar suggestions have been made with

respect to the learning of higher level cognitive skills (e.g., J. R. Anderson, Corbett, Koedinger, & Pelletier, 1995). Withholding feedback from the learner, it is thought, may force the individual to engage in deeper processing during learning and, thereby, improve later retention and generalization.

Although there is solid evidence that withholding feedback can have beneficial effects on delayed test performance in motor learning tasks (Tomlinson, 1972), studies involving acquisition of discrete verbal associations or factual information paint a fairly confusing picture. Several early studies suggested that feedback may have no effect on learning. In one such study, Schulz and Runquist (1960) trained subjects on paired associates, providing complete feedback on a predetermined fraction of the items (and the items that received feedback varied randomly from one presentation to the next, so that all items may have received feedback at some point). Subjects were tested 1 day after learning. There was no significant difference between the feedback conditions in the initial test performance on Day 2. However, training on Day 1 was to a criterion of one perfect recall of the whole list; thus, feedback was confounded with degree of practice, rendering the results inconclusive.

Two studies without this fatal confound also found no significant effect of feedback, however. R. C. Anderson, Kulhavy, and Andre (1972) had subjects read a programmed learning text (a text containing embedded questions pertaining to the material). Subjects were given feedback on all of the items or none of the items. There was no significant difference in performance on the final test, which was given shortly after the learning session. Although not significant, the difference did favor the 100% feedback condition. Krumboltz and Weisman (1962) also used a programmed learning text and provided various schedules of feedback, with some subjects experiencing no feedback on a fixed or variable one third or two thirds of items. The results revealed no significant differences on a session-final test. Thus, these two studies at least suggest that feedback does not make much difference.

Feedback After Errors or Correct Responses

The studies just described challenge the commonsense view that feedback is always helpful, but they do not provide strong support for the idea that withholding feedback is actually beneficial (e.g., Bjork, 1994; Rosenbaum et al., 2000; Schmidt & Bjork, 1992). After all, the retention intervals examined were no greater than 1 day.

These studies also fail to provide much clarity on the effect of feedback for another reason: because they examined only aggregate learning, without regard to whether the subject had made an error on a particular item. For various reasons, one might well expect that effects of feedback would differ depending on whether the response preceding the feedback was correct or incorrect. If one views learning from the perspective of contemporary neural-network learning models (Rizzuto & Kahana, 2001; Rosenblatt, 1959; Widrow & Hoff, 1960), feedback ought to be most critical after errors, because it is here that error-correction learning algorithms could be invoked (although some tuning may also be useful even after correct responses; cf. Mozer, Howe, & Pashler, 2004). The absence of an overall effect of feedback in the three studies described above does not necessarily refute the idea that feedback might be potentially helpful after errors, because in each of these studies, the error rate and the statistical power both appear quite low.

Specific evidence for the importance of distinguishing between feedback after errors and feedback after correct responses comes from a study by Guthrie (1971). Guthrie showed subjects sentences with one word missing; the correct response was the missing word. Feedback consisted of nothing, the

sentence with the missing word, or just the missing word. When subjects made an error, learning (as assessed on a test at the end of the learning session) was strongly facilitated by feedback, regardless of whether they were shown the sentence plus missing word or just the missing word. However, when subjects did not make an error on an item, feedback was not helpful.

Although Guthrie's (1971) data suggest that feedback can be important for discrete associative learning, they do not necessarily refute the idea that benefits of feedback may trade off against long-term retention (e.g., Bjork, 1994; Rosenbaum, et al., 2000; Schmidt & Bjork, 1992), because Guthrie's final test occurred within the same session. If there are benefits to withholding of feedback that only appear on subsequent days, these benefits would not have been apparent.

Another possibility is that after some partial learning has taken place, feedback consisting merely of labeling the preceding response "correct" or "wrong" might be quite useful. It is conceivable that this sort of partial feedback might prompt elaborate processing, producing benefits that would grow with delay of the test. For example, if a person makes an error of commission, producing the wrong response, and is told merely that the response is wrong, this might trigger him or her to engage in cognitive processes that would eliminate the competing response and allow the correct response to emerge (even, or especially, if that correct response is not provided at the time).

In summary, many divergent possibilities can be proposed for how feedback might operate, and what kind of feedback might be most effective, both for immediate performance and for memory assessed after some delay. Despite the theoretical and practical importance of this issue, existing data shed little light on these possibilities.

Present Approach

To provide a more fine-grained picture of how different kinds of feedback affect the acquisition and retention of associative information, we incorporated six design features in this study. First, we used a task with some face validity as a real educational problem (foreign-language vocabulary learning). Second, a nontrivial retention interval (1 week) was used to allow assessment of delayed learning. Third, we deliberately provided only a modest amount of instruction in the initial training session to ensure that both errors and correct responses would be plentiful during the learning sessions. The fourth feature, which was made possible by the third, was a fine-grained analysis of the course of learning at the item level, assessing final recall conditionalized on events occurring during earlier phases of the experiment. Fifth, subjects indicated their confidence in each response (a feature not included in any prior studies of this kind, to our knowledge). Sixth, we used web-based data collection to obtain a larger and more demographically diverse sample than is the norm in standard laboratory-based memory research.[1] This allows us to assess the generalizability of our findings across a broader range of ages and memory abilities than one normally encounters with college-student samples.

Method

Subjects. Subjects (N = 258) were recruited from a diverse online research subject panel. Subjects enrolled in this panel agreed to participate in a variety of behavioral science studies conducted by researchers in return for incentives such as enrollment in drawings for prizes.

1. Internet sampling methodology is increasingly common throughout psychology. Numerous lab/web replications have now been reported (e.g., Birnbaum, 1999; Krantz & Dalal, 2000; McGraw, Tew, & Williams, 2000; Reips, 2002), and in our own lab, we have found good agreement between results of lab- and web-based studies.

Materials and stimuli. A list of 20 words from the Luganda dialect and their English translations was assembled (available from the authors on request). This language was selected because the words tend to be fairly pronounceable but also unfamiliar to U.S. subjects (e.g., *leero* [today]). During the initial two presentations, the Luganda words were presented in black on a white background in a 20-column wide text box. The English translation was presented in an identical text box on the line immediately below the Luganda word. During the following test trials occurring during the learning session, the Luganda word was presented as follows: *"leero" means:*. The English word text box presented on the line below was empty, allowing the subject to type in his or her answer.

Design. The experiment was a between-subjects design with feedback condition as the sole independent variable. Subjects were randomly assigned to one of five feedback conditions. Following their response, subjects (1) immediately moved on to the next word being tested (0-s blank screen condition), (2) experienced a delay of 5 s (5-s blank screen condition), (3) saw the word *correct* or *incorrect* for 5 s (correct/incorrect condition), or (4) saw the correct answer for 5 s (correct-answer condition). An additional small amount of time, equated across conditions, separated the end of one trial and the presentation of the next word while the next browser page was loading. This additional time was less than 1 s in almost all cases. Finally, (5) an additional group of subjects experienced the initial exposures but no additional testing during Session 1 (not tested on Day 1 condition).

Procedure. Each subject participated in two sessions separated by a week, with some subjects completing the second session 1 day early or 1 day late. The first session was a training session. This consisted of two presentations of the entire list followed by two tests (conducted with procedures that depended on feedback conditions). Upon reading a brief description of the study and clicking the experiment link, subjects read a consent form, provided demographic information, and read instructions describing the procedure. In the initial presentation, all 20 pairs were presented successively for 6 s per pair, with a 2 s pause between pairs. This presentation was followed by a second learning presentation. Stimuli were presented in an independent random order during each learning presentation. Two learning tests followed (Tests 1 and 2). In each test, the stimuli were presented in an independent random order. On test trials, the Luganda word was presented with a response box below it, cuing the subject to type in the English word if they felt they might know the answer (the text box gave no cues for the number of letters to be typed). To respond, subjects could either check *I can't even guess* or type in an answer and indicate their confidence on a five-item scale ranging from *very low* to *very high*. (A reviewer pointed out that the use of a Likert-type scale limits our analysis to ordinal comparisons, whereas a scale using cardinal values, such as *60% likely to be correct*, might have given us both ordinal comparisons and evaluations of calibration and absolute accuracy.)

Subjects were free to take as long as needed to respond. After each response, the computer provided feedback according to the subject's condition. A response was considered correct if at least 70% of letters were correct, to allow for misspellings of the English word. This algorithm correctly distinguished correct and incorrect responses more than 99% of the time on the basis of double checking of 5% of answers by hand.

Twelve h prior to 7 days after first session completion (i.e., 6.5 days after Session 1), the server computer sent subjects an e-mail request to participate in Session 2. When the

subject clicked on a link in the e-mail, he or she was connected to the server, which presented the appropriate materials. Subjects were required to complete the Session 2 (Test Session) by 25 h after the 7-day time point. In this session, the subject was tested on all 20 items in a new random order (again, providing confidence for each response). There was no feedback given during the test session.

Results and Discussion

To assess performance, we first determined accuracy for each condition and test for each subject separately. These values were then averaged across subjects. Figure 1 shows the overall performance on Tests 1 and 2 (learning session) and final test (1 week later). Because subjects were assigned randomly to conditions that did not vary until after Test 1, differences in Test 1 can reflect only sampling error, and indeed, performance varied little between conditions.

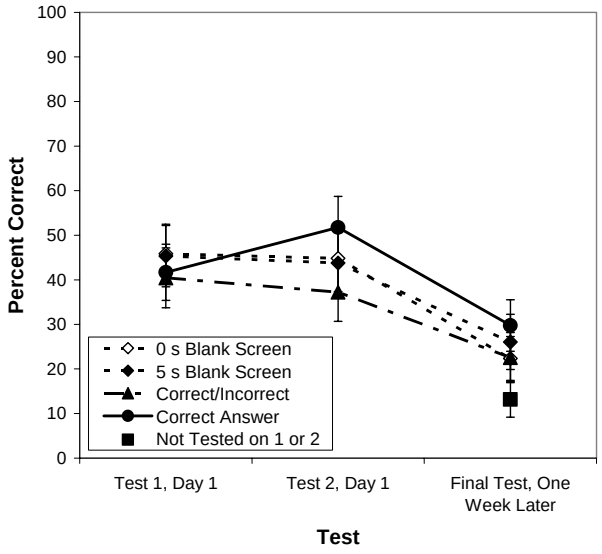

Figure 1. Accuracy in Experiment 1 for each type of feedback and for each test. Error bars represent standard errors.

The results beyond Test 1 show a clear pattern, with only the correct-answer feedback group showing improvement between Test 1 and Test 2. It is important to note that this group retained its advantage in the final test. A set of paired *t* tests confirmed that the correct-answer group showed improvement, as reflected in a difference between Test 1 and Test 2 for the correct-answer condition, $t(59) = 5.2$, $p < .01$. Zero- and 5-s blank screen conditions did not show learning, $t(52) = 1.1$, $p = .278$, and $t(47) = 1.6$, $p = .113$, respectively, whereas the correct/incorrect condition actually showed a small decrease in recall, rather than an increase, between Tests 1 and 2, $t(44) = 2.6$, $p < .05$.

For a more fine-grained analysis of the effects of feedback, we examined performance on Test 2 and the final test conditionalized on performance on Test 1. We determined conditional accuracy for each cell for each subject. In Figure 2, Panel A shows performance on trials in which the correct response was made on Test 1, whereas Panels B and C show performance on trials in which Test 1 elicited no response (Panel B) or an incorrect response (Panel C). The first thing one notices is dramatically better overall performance in Panel A where the correct response was made on Test 1. This is unsurprising and, presumably, reflects differences in item difficulty as well as amount of initial learning. The second finding, visible in Panel A, is that when Test 1 was correct, feedback condition made little difference. To show this, we conducted a mixed-model analysis of variance with test (Test 2 vs. final) and feedback condition (0- vs. 5-s blank screen, correct incorrect, correct answer) as factors. We found a main effect of test, $F(1, 186) = 209.8$, $p < .01$, but no main effects or interactions involving feedback condition (all ps $> .05$). Independent samples *t* tests of final test data showed no significant differences between any feedback conditions (all ps$>.05$).

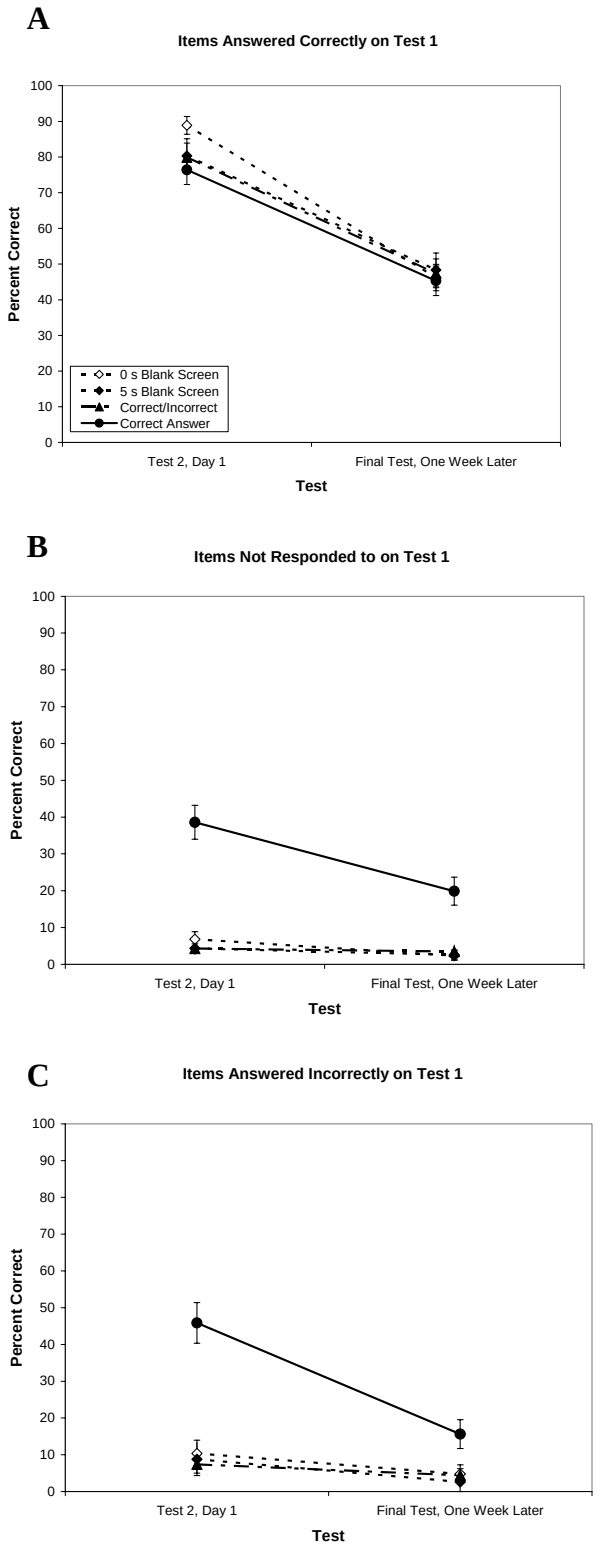

Figure 2. Percent correct answers on later tests, when the subject provided the correct response (Panel A), no response (Panel B), or the wrong response (Panel C) on the first test, by feedback condition, for Experiment 1. Error bars represent standard errors.

By contrast, in Panels B and C depicting performance after errors of omission and commission on Test 1, one sees a dramatic effect of feedback. Independent samples *t* tests confirmed that the correct-answer feedback condition showed better final-test performance than did any other condition (all pairwise comparisons of correct-answer vs. other feedback conditions, $p < .05$; all other ps $> .05$). As noted earlier, it has recently been suggested that feedback may impair longer term retention, but as seen in Panels B and C, supplying the correct answer after an incorrect response increased 1-week retention by 494% as compared with the no-feedback condition.

Table 1 shows the relation between subjects' confidence in their responses on the final test and their performance on this test (we computed this by averaging the accuracy of all the items to which a given subject assigned a particular confidence level and then averaging these values across subjects). For levels of confidence ranging from *very low* to *very high*, the overall correct response rates were 15%, 37%, 61%, 84%, and 90%, confirming that (not surprisingly) confidence was very closely related to accuracy.

Table 1

Relationship between Confidence on the Final Test Answer and Final Test Percent Recall, for each Feedback Condition

	Confidence in Final Test Answer				
Feedback Condition	1 (Very Low)	2	3	4	5 (Very High)
0 s Blank Screen	10.7	47.5	69.4	82.9	87.3
5 s Blank Screen	11.5	39.2	56.1	82.1	95.7
Correct / Incorrect	25.5	30.5	52.4	84.6	86.1
Correct Answer	23.1	42.5	57.9	87.7	94.7
Not Tested on 1 or 2	5.4	26.3	69.8	81.3	83.8

Table 2 shows percent correct on the final test, partitioned by feedback confidence and each subject's confidence in his or her Test 1 answer for items eliciting a correct responses on Test 1. Many subjects made no wrong answers when they were very confident in being correct on Test 1, and some subjects never used medium confidence values. Thus, the number of subjects contributing to each mean varied. When subjects responded correctly (Table 2), final test accuracy (correct vs. incorrect) was closely and positively related to Test 1 confidence. It is of interest that the three conditions in which correct-answer feedback was withheld did not show any gross impairment in performance at low confidence levels. One might have thought that when subjects responded correctly, doubted they were correct, and were given no alternative response, they might effectively weaken the link to the response they had just made (treating their own skepticism much as they would treat feedback from the experimenter indicating that they had made an error). Although the data may not completely rule this out, they offer no encouragement for it.

Table 2

Relationship between Confidence on the Test 1 Answer and Final Test Percent Recall, for each Feedback Condition, for Items that Were Correct on Test 1. Standard Errors Are Shown in Parentheses

	Confidence in Test 1 Answer				
Feedback Condition	1 (Very Low)	2	3	4	5 (Very High)
0 s Blank Screen	4.0 (4.0)	30.8 (13.3)	31.7 (8.9)	34.8 (8.6)	54.8 (4.1)
5 s Blank Screen	28.6 (12.5)	27.3 (14.1)	42.1 (11.0)	33.3 (8.2)	59.7 (4.6)
Correct / Incorrect	22.2 (16.5)	31.8 (13.9)	41.1 (10.1)	49.2 (10.3)	53.5 (4.8)
Correct Answer	23.3 (13.2)	26.3 (9.6)	35.2 (8.2)	46.1 (6.5)	51.0 (4.9)

The basic pattern of results described above was confirmed in another, quite similar, online experiment that we carried out teaching subjects obscure facts rather than foreign language vocabulary, and using a within-subject instead of a between-subjects manipulation of the type of feedback. Thus, we suspect the results will prove quite general.

Practical and Theoretical Implications

Although the importance of feedback after erroneous responding seems rather commonsensical, as described in the introduction to this article, it has been challenged by recent commentaries proposing that feedback facilitates performance in training at the expense of actual learning (e.g., Bjork, 1994; Rosenbaum et al., 2000; Schmidt & Bjork, 1992). On the basis of the present results, one would suspect that although withholding feedback may be useful in certain motor learning contexts, it is likely to be counterproductive in discrete verbal learning tasks requiring explicit cued recall. Naturally, this does not undermine the more general point that the way to improve enduring memory is not necessarily through procedures that improve performance during a learning session (Schmidt & Bjork, 1992).

However, the results indicate that when the learner makes a correct response, feedback makes little difference for what can be remembered 1 week later. Given these findings, along with the lack of any benefit from the 5-s pause condition in Experiment 1, a reasonable strategy for computer-aided instruction would seem to be this: Whenever the subject makes an error, provide feedback and time to process the feedback, but when the subject responds correctly, proceed to the next trial without delay (which would, in any case, allow the learner to infer that his or her past response was correct; Nelson, 1971).

The Skinnerian reinforcement-based perspective seems to shed little light on the kinds of learning studied here. Telling subjects

they were right after they made a correct response produced no detectable improvement in either immediate performance or learning as assessed in a delayed test. This is consistent with early results indicating that the critical factor in reinforcement is whether the subject can infer which response would be regarded as correct in the future, not a putative stamping in of a behavior by the satisfaction that follows from feedback (Buchwald, 1969; Nelson, 1971, 1977; for discussion, cf. Kulhavy, 1977; McKeachie, 1974). Although the Skinnerian perspective seems far off the mark, the error-correction learning framework (Mozer et al., 2004; Rizzuto & Kahana, 2001; Rosenblatt, 1959; Widrow & Hoff, 1960) seems quite congenial to the overall pattern of results described here. It implies that underlying cognitive representations are tuned up on the basis of perceived mismatches between the response to a particular cue that the system is inclined to produce and the response that it should have produced.

References

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *Journal of the Learning Sciences, 4*, 167–207.

Anderson, R. C., Kulhavy, R. W., & Andre, T. (1972). Conditions under which feedback facilitates learning from programmed lessons. *Journal of Educational Psychology, 63*, 186–188.

Birnbaum, M. (1999). Testing critical properties of decision making on the internet. *Psychological Science, 10*, 399–407.

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

Buchwald, A. M. (1969). Effects of "right" and "wrong" on subsequent behavior: A new interpretation. *Psychological Review, 76*, 132–143.

Butterfield, B., and Metcalfe, J. (2001). Errors committed with high confidence are hypercorrected. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*, 1491–1494.

Gagné, R. M., Briggs, L. E., & Wager, W. W. (1992). *Principles of Instructional Design, 4th Edition.* Fort Worth, TX: Harcourt Brace Jovanovich.

Guthrie, J. T. (1971). Feedback and sentence learning. *Journal of Verbal Learning and Verbal Behavior, 10*, 23–28.

Krantz, J. H., and Dalal, R. (2000). Validity of web-based psychological research. In M. Birnbaum (Ed.), *Psychological experiments on the internet*. San Diego, CA: Academic Press.

Krumboltz, J. D., & Weisman, R. G. (1962). The effect of intermittent confirmation in programmed instruction. *Journal of Educational Psychology, 53*, 250–253.

Kulhavy, R. W. (1977). Feedback in written instruction. *Review of Educational Research, 47*, 211–232.

McGraw, K. O., Tew, M. D., & Williams, J. E. (2000). The integrity of web-delivered experiments: Can you trust the data? *Psychological Science, 11*, 502–506.

Mozer, M. C., Howe, M., & Pashler, H. (2004). Using testing to enhance learning: A comparison of two hypotheses.

Proceedings of the Cognitive Science Society.

Nelson, T. O. (1971). Extinction, delay, and partial-reinforcement effects in paired-associate learning. *Cognitive Psychology*, *2*, 212–228.

Nelson, T. O. (1977). Reinforcement and human learning. In W. K. Estes (Ed.), *The Psychology of learning and memory* (pp. 207–246). Hillsdale, NJ: Erlbaum.

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.

Reips, U.-D. (2002). Standards for internet experimenting. *Experimental Psychology*, *49*, 243–256.

Rizzuto, D. S., & Kahana, M. J. (2001). An autoassociative neural network model of paired-associate learning. *Neural Networks*, *13*, 2075–2092.

Rosenbaum, D. A., Carlson, R. A., & Gilmore, R. O. (2000). Acquisition of intellectual and perceptual-motor skills. *Annual Review of Psychology*, *52*, 453–470.

Rosenblatt, F. (1959). *Principles of neurodynamics*. New York: Spartan Books.

Schmidt, R. A., & Bjork, R. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science*, *3*, 207–217.

Schulz, R. W., & Runquist, W. N. (1960). Learning and retention of paired adjectives as a function of percentage occurrence of response members. *Journal of Experimental Psychology*, *59*, 409–413.

Tomlinson, R. W. (1972). Control impedance and precision of feedback as parameters in sensori-motor learning. *Ergonomics*, *15*, 33–47.

Widrow, B., & Hoff, M. E. (1960). Adaptive switching circuits. In J. A. Anderson & E. Rosenfield (Eds.), *Neurocomputing: Foundations of research* (pp. 126–134). Cambridge, MA: MIT Press.

Harold Pashler, Nicholas J. Cepeda, and John T. Wixted, Department of Psychology, University of California, San Diego; Doug Rohrer, Department of Psychology, University of South Florida.

This work was supported by the Institute of Education Sciences (U.S. Department of Education, Grant R305H020061) and the National Institute of Mental Health (Grants R01 MH61549 and R01 MH45584). Ed Vul provided expert assistance with web programming and page design. David Perlmutter suggested the use of the Luganda language.