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

0278-7393/05/$12.00 DOI: 10.1037/0278-7393.31.1.3

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

Some researchers have suggested that although feedback may enhance performance during associative learning, it does so at the expense of later retention. To examine 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. A final test followed after 1 week. 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.

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

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.

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

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 hr 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.

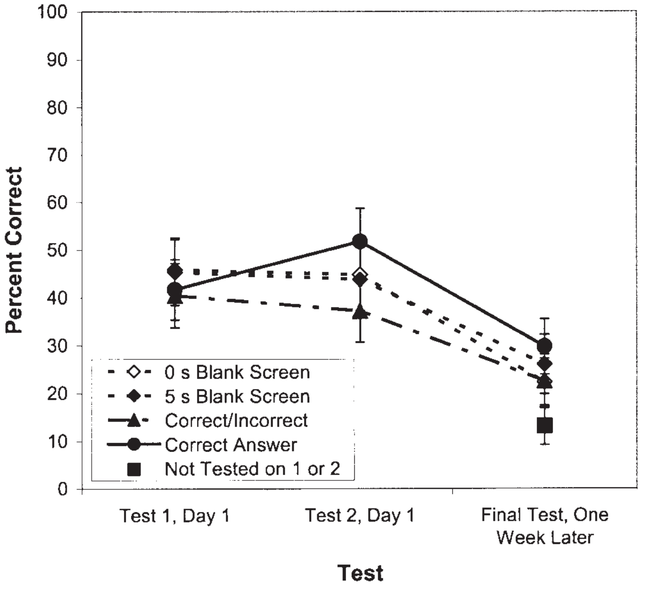

Figure 1. Accuracy in Experiment 1 for each type of feedback and for each test. Error bars represent standard errors.

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 $> .05$).

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