CONTEMPORARY EDUCATIONAL PSYCHOLOGY 16, 192–201 (1991)

BRIEF RESEARCH REPORT

Tests and Test Feedback as Learning Sources

MARK A. MCDANIEL
Purdue University

AND

RONALD P. FISHER
Florida International University

Two experiments examined the influence of test taking and feedback in promoting learning. Participants were shown a list of trivia facts during an incidental learning task. Some facts were later tested (plus feedback provided), whereas other facts were not presented for further processing. Tested facts were better recalled on a final criterion test than untested facts, showing the beneficial effects of testing. Tested facts were also better recalled than facts that were presented for additional study (Experiment 1). Although testing plus feedback enhanced learning, there were no effects of whether the participants were required simply to repeat the feedback or elaborate it. © 1991 Academic Press, Inc.

This report focuses on the influence of test taking and feedback in promoting learning. Two issues were examined. Experiment 1 investigated the value of test taking (plus feedback), compared to additional study time, on a later criterion test. A second issue, investigated in both experiments, concerns the extent to which the kind of processing applied to the test feedback influences performance on a later criterion test. This is an important issue because research on the mnemonic effects of test feedback has in large part not focused on the influence of behaviors that occur once feedback is presented (Kulhavy & Stock, 1989). Furthermore, at least with immediate feedback, forcing the learner to process and attend to the feedback can produce substantial gains on a later criterion test (Phye & Andre, 1989). Phye and Andre concluded that future research

We thank Jill Schaff for assistance in material preparation and subject testing, and also Brian Lyman and Cheryl Walker for help with subject testing. We are also grateful for comments provided by three anonymous reviewers on an earlier version of the paper. Preparation of this article was supported in part by the National Institute of Child Health and Human Development Grant No. HD23984. Requests for reprints should be sent to Mark A. McDaniel, Department of Psychological Sciences, Purdue University, West Lafayette, IN 47907.

0361-476X/91 $3.00

would profit by exploring a number of alternative procedures for focusing attention on the feedback. As a modest beginning in that direction, we examined two procedures for inducing learners to process feedback.

EXPERIMENT 1

The stimuli were little-known facts culled from the Nelson and Narens (1980) norms and the *Trivial Pursuit* game. Research with word lists has shown that an intervening test can enhance learning more so than a second study session. Performance is further boosted when test feedback is provided (see Kulhavy & Stock, 1989). On the basis of these findings we hypothesized that subjects who were given an initial test with feedback on the set of encoded facts would perform better on a criterion test than subjects who were not tested initially, but instead received an additional study session.

Most important for the present purposes, we manipulated the kind of processing that subjects were instructed to perform on the feedback provided after testing or on the second presentation of the facts. Some subjects were instructed to elaborate the target facts (presented as test feedback to some subjects and as a second presentation to other subjects) by providing some reason why each fact might be true. Other subjects were instructed to rote rehearse the target facts by repeating each fact out loud. On the basis of previous research and theory, we posited that the elaborative processing conditions should promote better performance on the final criterion test than the rote rehearsal conditions.

Method

Materials. Fifty-four facts were assembled from two sources. Nine facts were selected from the Nelson and Narens (1980) general information norms. These facts were known by less than 7.5% of the undergraduates tested in Nelson and Narens. Many of the Nelson and Narens corpus involved a fact about a particular person. To develop a stimulus set that included a wider range of factual material, we also selected facts from a popular trivia game (the *Genus* edition of *Trivial Pursuit*). We selected 45 additional facts that we judged to be relatively unknown by current undergraduates (confirmed with an informal sampling of undergraduate students at Notre Dame who were familiar with the *Trivial Pursuit* game). The complete stimulus set is provided in the Appendix.

Design. The design was a $2 \times 2 \times 2$ mixed factorial with two between-subjects factors and one within-subjects factor. For one between-subjects variable, half of the subjects received an additional study session and half received a test plus feedback. The other between-subjects variable involved the type of processing (rote or elaborative rehearsal) that subjects were instructed to perform during the additional study session or on the feedback given in the test session. The within-subjects variable was whether a fact was presented for additional study/immediate testing. For each subject, 18 of the facts seen during an initial

incidental encoding phase were *not* included for additional study or testing, and 36 facts were presented for either additional study or testing. The particular facts used in each of these sets were counterbalanced across subjects.

Subjects. Ninety subjects were tested: 60 introductory psychology students enrolled at the University of Notre Dame and 30 introductory psychology students enrolled at Purdue University. All subjects participated either for extra credit (Notre Dame) or in partial fulfillment of a course requirement (Purdue). Random assignment of subjects to groups produced 22 subjects each in the additional study and initial test groups that performed rote rehearsal and 23 subjects each in the additional study and initial test groups that performed elaborative rehearsal. Each group consisted of approximately equal numbers of students from each university sampled.

Procedure. Subjects initially encoded the entire set of target facts by rating the comprehensibility of each factual statement. This was an incidental encoding task as no mention was made of subsequent testing of the facts. Each of the 54 facts appeared on a CRT screen for 6 s, with the order of presentation randomized separately for each subject. After each 6-s presentation, the fact was replaced by a comprehensibility rating scale, with the endpoints anchored by 1 (indicating that the fact was "very clear") and 5 (indicating that the fact was "very unclear"). The subjects entered their rating by depressing the appropriate number on the keyboard. As soon as the rating was entered, the next fact appeared on the screen. An example statement and rating were provided at the outset to familiarize subjects with the procedure.

Once subjects finished the initial encoding task, the experimenter booted-up the software appropriate to the next phase of the experiment. During this 2- to 3-min interval, subjects were occupied with filling out a "need-for-cognition" self rating. Subjects worked on this self-rating task only until the experimenter had finished setting up the next phase of the experiment.

In this next phase, the additional study groups were represented with 36 of the initial 54 facts. Each fact was presented on the CRT for 5 s after which the screen was blank for 5 s. The rote rehearsal group was instructed to start repeating the fact aloud as soon as it appeared and to continue doing so until a tone sounded (10 s later), signaling the onset of another fact. The elaborative rehearsal group was instructed to "state out loud a plausible reason why this sentence is true." For the test-plus-feedback groups, subjects were presented with one question about each of 36 facts from the original set of 54. Each question appeared on the screen for 5 s, during which time subjects tried to answer the question. After 5 s the question was replaced with its answer. The rote rehearsal subjects were given 10 s to repeat aloud the answer as many times as they could. (The answer was placed in the frame of the question, so that these answers repeated the complete factual statement.) The elaborative rehearsal subjects were instructed to use the 10 s to "give a plausible reason why the answer is true." All subjects' verbal responses were tape recorded. After finishing this phase of the experiment, subjects were dismissed and instructed to return to the laboratory the next day.

In the final session subjects were given the criterion test, which consisted of 54 questions of the 54 target facts originally encoded Across subjects, six question orders were used, and approximately equal numbers of subjects were assigned to each particular ordering. One additional feature of the criterion test bears mention. There are various kinds of information embedded in a factual statement that can be targeted by a question. For example, for the fact, "The USSR agreed to enter into war against Japan at the Tehran Conference," one can ask either "At the Tehran Conference the USSR agreed to enter into war against which country?" or "At what conference did the USSR agree to enter war against Japan?" There were two questions for each fact so that the results would not be subject to possible biases due to the specific form of the question. Different versions of the criterion test were con-

structed so that across subjects, each question format was used equally often.[1] Subjects were allowed 10 min to write their responses to the criterion-test questions.

Results and Discussion

Initial recall. The rejection level was set at .05 for the statistical tests reported throughout this article. Subjects provided the correct answer for .49 of the initial test questions. The values for the rote rehearsal group (.51) and the elaborative rehearsal group (.45) did not differ significantly, as would be expected [$F(1,43) = 1.43$].

Criterion test recall. Table 1 provides the mean proportion of facts recalled on the criterion test as a function of format of postencoding experience (additional study versus testing-plus-feedback), type of processing (rehearsal, elaboration) during the postencoding session, and whether or not an item was presented during the postencoding session. These data were analyzed with a three-factor mixed analysis of variance (ANOVA), in which the first two factors were between-subjects variables and the third factor was a within-subjects variable.

Overall, final recall was better for postencoded items (presented for additional study or immediate testing) than for items that were not presented after the initial encoding, $F(1,86) = 277.55$, $MSe = .01$. This variable interacted with the format of postencoding experience, $F(1,86) = 5.87$, $MSe = .01$. The interaction reflects the fact that there were no differences between additional study and testing-plus-feedback for the nonpresented items; however, testing-plus-feedback was more beneficial than additional study for the postencoded items.

Type of processing during the postencoding session, whether rehearsal or elaboration, did not significantly affect performance ($F = 2.65$, for the main effect). Type of processing did not interact with whether or not items were presented for postencoding processing ($F = 1.14$), nor did it interact with the format of the postencoding experience (study or test) ($F = 2.12$). The three-way interaction was also not significant ($F < 1$).

Thus, the major findings were (1) testing with feedback produced more learning than additional study and (2) instructions to elaborate, for either an additional study trial or for test feedback, did not enhance learning (as measured by recall) relative to rote rehearsal instructions. These results will be discussed under General Discussion.

[1] Note that for the test-plus-feedback groups, some of the final-test questions were in exactly the same format as those that appeared on the immediate test, whereas others were in an alternative format. The mix of "same format" and "different format" questions was haphazard, although across test-plus-feedback subjects each factual statement was tested equally often with same- and different-format questions. In Experiment 2 this factor was arranged so that it could be included as an independent variable.

TABLE 1
PROPORTION CORRECT ON THE CRITERION TEST IN EXPERIMENT 1

	Postencoding task			
	Additional study		Initial test plus feedback	
Item	Rehearsal	Elaboration	Rehearsal	Elaboration
Postencoded items	.62 (.15)	.64 (.10)	.74 (.14)	.65 (.14)
Not postencoded items	.42 (.17)	.39 (.16)	.44 (.16)	.34 (.22)
Difference	.20	.25	.30	.31
N per group	22	23	22	23

Note. Standard deviations are in parentheses.

EXPERIMENT 2

In Experiment 1, taking a test (plus feedback) was superior to having an additional study session. Therefore, in Experiment 2 we focused on two procedures for inducing attention to test feedback. Specifically, we were interested in why there were no differences between the elaborative and rote repetition procedures. We thought it possible that instructions to process test feedback elaboratively could significantly influence subsequent recollection if subjects had enough time to try to generate elaborations for most or all of the target items. Accordingly, we increased the time that subjects were given to process the feedback from 10 s (in Experiment 1) to 18 s.

Second, we speculated that if the criterion test were more difficult, elaboration of feedback might be more effective than rehearsing feedback. Accordingly in Experiment 2 we increased the interval between the initial and criterion tests from 24 h (used in Experiment 1) to 48 h. We also manipulated the similarity between the questions posed during initial and final testing. For half of the tested facts, the question was exactly the *same* for both the initial and criterion tests. For the other half of the tested facts, the answer on the criterion test involved a *different* aspect of the fact than that probed on initial testing. For example, for the fact, "In Moslem countries white is the mourning color," the initial question was "What's the mourning color in Moslem countries?" and the criterion question was "What does the color white symbolize in Moslem countries?" Presumably, the different-format test condition should be more difficult than the same-format test condition, thereby allowing the possibility that elaborative processing of feedback may become more potent than rote rehearsal of feedback in at least the different-format condition.

Method

Subjects and design. Students in introductory psychology classes at Purdue University participated in partial fulfillment of a course requirement. The design was a 2 × 3 mixed factorial. The between-subjects factor involved how subjects were instructed to process the feedback for the initial test (rehearse vs. elaborate); 12 subjects were randomly assigned to each of these two groups. The within-subjects factor, the relation between the criterion and initial test questions, had three levels. Some of the facts probed on the criterion test had not been tested initially, some were tested with the *same* question frame as was used on the initial test, and some were tested with a *different* question frame than was used on the initial test.

Procedure. The set of facts from Experiment 1 was used. Subjects initially encoded the facts as in Experiment 1. There were two changes in the initial test procedure from that in Experiment 1. The presentation rates were extended so that the question was presented for 7 s and the feedback was presented for 18 s. Also, subjects wrote their answers to the questions during the 7-s questioning period (as in Experiment 1, subjects still verbally responded to the feedback). The criterion test was the same as in Experiment 1, except that it was administered 2 days after initial testing.

Results and Discussion

Initial recall. Overall, subjects provided the correct answer for .49 of the initial test questions. A 2 × 2 mixed ANOVA (with type of feedback processing and similarity of initial and criterion test questions) indicated that the values for each of the experimental cells did not differ significantly from one another. This would be expected given that the independent variables should not have influenced initial test performance.

Criterion test recall. We first examined the proportion recalled on the criterion test as a function of how the information was tested on the initial test (not tested, tested with same question as on the criterion test, tested with a differently framed question) and how subjects were instructed to process the initial test feedback (rehearse, elaborate). A two-factor mixed ANOVA indicated that instructions to elaborate feedback on the initial test did not produce significantly better criterion test recall than instructions to rehearse feedback [$F(1,22) = 1.21$] (see Table 2 for means). There was a significant main effect of how information was tested initially, $F(2,44) = 33.68$, $MSe = .03$. Examination of Table 2 indicates that this effect was due to tested items being better recalled than nontested items. A post-hoc comparison showed that for tested items, same- and different-format items did not significantly differ.

The possible effects of the format of the criterion test question (same or different from the initial test question) on final performance were explored further by examining final recall of questions as a function of whether they were answered correctly on the initial test. A three-factor mixed ANOVA was conducted, with format of criterion test question (same, different) and performance on initial question (correct, incorrect) as within-subjects variables and type of feedback processing (rehearsal,

TABLE 2
PROPORTIONAL CORRECT ON THE CRITERION TEST IN EXPERIMENT 2

Immediate test question	Feedback processing	
	Rehearsal	Elaboration
None	.25 (.19)	.40 (.26)
Same-format	.70 (.15)	.72 (.18)
Immediate correct	.91 (.12)	.95 (.08)
Immediate incorrect	.44 (.18)	.44 (.27)
Different-format	.63 (.24)	.67 (.20)
Immediate correct	.79 (.25)	.75 (.25)
Immediate incorrect	.44 (.35)	.40 (.28)
N per group	12	12

Note. Standard deviations are in parentheses.

elaboration) as a between-subjects variable. This analysis indicated that those items answered correctly on the initial test were significantly better recalled on the criterion test than those items not answered correctly initially, $F(1,22) = 81.39$, $MSe = .05$. Criterion test question format did not interact significantly with the correctness of the initial answer, $F(1,22) = 2.31$. However, a comparison of the effects of question format for items correctly answered on the initial test showed that performance on the criterion test with a same-format question was better than performance with a different-format question (.93 vs. .77), $F(1,22) = 6.15$, $MSe = .05$. For items answered incorrectly on the initial test, question type was not influential ($F < 1$).

GENERAL DISCUSSION

Two focal results emerged from this study. First, elaborative processing did not enhance later recall compared to rote rehearsal. Consistent with Walker (1986), the present findings suggest that retrieval of arbitrary associations is not necessarily enhanced by trying to elaborate on these associations relative to simple repetition of the associations. It would be premature, however, to (1) generalize this specific conclusion to more educationally representative situations or (2) conclude that how the learner processes feedback has little effect on the benefit of feedback so long as the learner attends to the feedback. In a metaanalysis of feedback effects, Kulik and Kulik (1988) found different patterns for "applied" studies (using actual classroom quizzes and real learning materials) compared to laboratory studies. More particularly, presenting lists of unrelated facts, as was done in both the present study and Walker's study, may involve different learning processes than presenting facts within an integrated lesson. Further, in an educational setting the dynamics of feedback presentation would likely be different than those used here. For

instance, in the classroom, feedback would ordinarily be provided after the test rather than immediately after each item, and learners would not be as limited in the time available to process the feedback as they were in the present study. These limitations notwithstanding, the current study extends previous work on feedback processing by focusing on recall rather than multiple-choice tests (cf. Kulhavy & Stock, 1989) and by including an initial encoding of target content prior to the first test (e.g., in Phye & Andre, 1989, there was no instruction on the content targeted by the tests).

Second, testing with feedback increased learning more than providing an additional study trial. Given our failure to find effects of elaborative processing, the positive effect of test taking with feedback is even more important. In the present experiment, not only did test taking improve recall, but it did so even more than did an additional study session, the traditional medium for imparting knowledge. From an educational viewpoint, then, this procedure may contain considerable potential that is currently underutilized.

As a final note, recall that the value of test taking was a joint function of the success of retrieval on the initial test and the format of the question on the final criterion test (Experiment 2). The finding that initially retrieved items were recalled better on the delayed test than items not initially retrieved may not be very telling, because of possible item selection artifacts. Those items that were answered correctly on the initial test may have been easier than those not answered correctly initially. However, of those items that were retrieved successfully initially, criterion performance was most enhanced when the final question was in the same form as the initial question. This finding converges with that of recent research using both text-like and more traditional laboratory materials, suggesting that the mnemonic benefits of retrieval are tied to the cue that initiates later recall.

APPENDIX
STIMULI FOR EXPERIMENTS 1 AND 2

1. The oval office has been called the innermost sanctuary of American power.
2. The largest city in the communist world is Shanghai.
3. God called heaven the firmament.
4. Pocahontas is buried along the Thames river.
5. Hebrew and Aramaic were the two original languages of the Old Testament.
6. The Smithsonian Institution was once called America's Attic.
7. Vampire bats usually attack sleeping humans' toes.

8. The turkey is the dumbest domesticated animal.
9. One-ninth of an iceberg shows above water.
10. Apricots were the golden apples of Greek mythology.
11. A bungey launch or a car-tow can be used on a glider.
12. There are two versions of the Ten Commandments in the Bible.
13. Cyprus is the only Mediterranean country to display its map on its flag.
14. The forget-me-not is the state flower of Alaska.
15. Barry Goldwater declared, "Extremism in the defense of liberty is no vice."
16. Denmark sold the Virgin Islands to the United States.
17. Prudence, justice, temperance, and fortitude are the four cardinal virtues.
18. Boris Onishchenko was caught cheating in the sport of fencing at the Montreal Olympics.
19. There are five varieties of twins.
20. Statistically 10 is the safest age of life.
21. Cyclamates got the ax in 1969.
22. The U.S.S.R. agreed to enter into war against Japan at the Tehran Conference.
23. A row of crows is called a murder.
24. One percent of the earth's water is drinkable.
25. The principle of conservation of energy makes a perpetual motion machine an impossibility.
26. At the Casablanca Conference F.D.R. and Churchill announced their policy of unconditional surrender.
27. The Everly Brothers' song "Wake up Little Susie" was banned in Boston.
28. The stars and stripes flies over Wake Island.
29. Intourist is the name of the Soviet Union's state run travel agency.
30. The machine gun was the favorite weapon of George Kelly and Kate Barker.
31. The Canary Islands were named for dogs.
32. The song "Fire and Rain" put James Taylor in the limelight.
33. Lyndon B. Johnson's first presidential order was, "Let's get this god-damned thing airborne."
34. There are two talmuds.
35. The most popular contact lens color is blue.
36. Christ's zodiacal sign was Capricorn.
37. Rhubarb and asparagus are the only perennial vegetables.
38. The 737 Boeing jet is nicknamed Fat Albert.
39. California grants the most fishing licenses in the United States.
40. The queen was in the parlor eating bread and honey.

41. Sir Freddie Laker's life story is entitled, "Fly Me, I'm Freddie."
42. In Moslem countries white is the mourning color.
43. Idi Amin seized power from Milton Obote.
44. Pills are the dolls in Jacqueline Susann's "Valley of the Dolls."
45. Israel offered Albert Einstein its presidency.
46. Andersonville was the largest Confederate military prison during the Civil War.[2]
47. Bing Crosby's theme song was "When the Blue of the Night Meets the Gold of the Day."
48. The number two wood in golf is called the brassie.
49. Lon Chaney was known as "the man of a thousand faces."
50. Bagdad is the capital of Iraq.
51. Charlemagne was the first ruler of the Holy Roman Empire.
52. Angel Falls is located in Venezuela.
53. Batista is the Cuban leader that Castro overthrew.
54. Trevi is a fountain in Rome into which coins are thrown for good luck.

REFERENCES

KULHAVY, R. W., & STOCK, W. A. (1989). Feedback in written instruction: The place of response certitude. *Educational Psychology Review,* **1,** 279–308.

KULIK, J. A., & KULIK, C. C. (1988). Timing of feedback and verbal learning. *Review of Educational Research,* **58,** 79–97.

NELSON, T. O., & NARENS, L. (1980). Norms of 300 general-information questions: Accuracy of recall, latency of recall, and feeling-of-knowing ratings. *Journal of Verbal Learning and Verbal Behavior,* **19,** 338–368.

PHYE, G. D., & ANDRE, T. (1989). Delayed retention effect: Attention, Perseveration, or both? *Contemporary Educational Psychology,* **14,** 173–185.

WALKER, N. (1986). Direct retrieval from elaborated memory traces. *Memory & Cognition,* **14,** 321–328.

[2] This stimulus and the remaining eight stimuli are from Nelson and Narens (1980).