The Psychological Record, 2002, 52, 187-201

IMMEDIATE FEEDBACK ASSESSMENT TECHNIQUE PROMOTES LEARNING AND CORRECTS INACCURATE FIRST RESPONSES

MICHAEL L. EPSTEIN, AMBER D. LAZARUS, TAMMY B. CALVANO, KELLY A. MATTHEWS, RACHEL A. HENDEL, BETH B. EPSTEIN, and GARY M. BROSVIC
Rider University

Multiple-choice testing procedures that do not provide corrective feedback facilitate neither learning nor retention. In Studies 1 and 2, the performance of participants evaluated with the Immediate Feedback Assessment Technique (IF AT), a testing method providing immediate feedback and enabling participants to answer until correct, was compared to that of participants responding to identical tests with Scantron answer sheets. Performance on initial tests did not differ, but when retested after delays of 1 day or 1 week, participants evaluated with the IF AT demonstrated higher scores and correctly answered more questions that had been initially answered incorrectly than did participants evaluated with Scantron forms. In Study 3, immediate feedback and answering until correct was available to all participants using either the IF AT or a computerized testing system on initial tests, with the final test completed by all participants using Scantron forms. Participants initially evaluated with the IF AT demonstrated increased retention and correctly responded to more items that had initially been answered incorrectly. Active involvement in the assessment process plays a crucial role in the acquisition of information, the incorporation of accurate information into cognitive processing mechanisms, and the retrieval of correct answers during retention tests. Results of Studies 1-3 converge to indicate that the IF AT method actively engages learners in the discovery process and that this engagement promotes retention and the correction of initially inaccurate response strategies.

Testing and assessment are integral to the educational process. When university or college education takes place as tutorials or in classrooms with a small number of participants, essay examinations are preferred, as they are relatively easy to construct, they allow the instructor to assess the depth and breadth of participant understanding, and they

Correspondence concerning this article should be addressed to Michael L. Epstein, Department of Psychology, Rider University, 2083 Lawrenceville Road, Lawrenceville, New Jersey, 08648. (E-mail: Epstein@ Rider.edu).

enable the instructor to allocate partial credit for proximate knowledge. There are, however, significant drawbacks to the essay format, including subjectivity in scoring, variation in the quality and quantity of feedback within and between evaluators, and the substantial investment of time, energy, and attention to score. The administration of essay questions in large classes typically lengthens the amount of time between the completion and the return of examinations, and in many cases, decreases the amount of corrective information that can be supplied.

One solution to several of these drawbacks is the use of the multiple-choice test format. Mislevy (1991) discusses the origins and explosive growth in multiple-choice testing since World War 1. Educators teaching classes with small and large enrollments found that multiple-choice tests were easy to score, were reliable, minimized subjectivity, and could often be returned at the next class meeting. The advent of computerized test banks has made test construction a simple process. Although, in many circumstances, multiple-choice tests are more appropriate than essay examinations, they too have drawbacks. Multiple-choice tests tend to be difficult to construct in the absence of a publisher-supplied test bank, and given the necessity of a single best answer, they are not as sensitive to proximate knowledge as the essay format. Also, a multiple-choice question is often related either to an earlier or to a subsequent test question, and thus an incorrect response on one item will likely be associated with a similar error on the related item—a type of "double jeopardy."

Among the more substantive drawbacks of both test formats are the failure to facilitate learning during the test-taking process and the return of either instructor- or machine-scored tests without information to correct inaccurate responding, an essential feature of the learning process. Despite almost a century of research, there is little consensus either about the mechanisms by which feedback affects learning or about the efficacy of feedback (e.g., Kluger & DeNisi, 1998). Delays as short as several seconds have been reported to adversely affect the learning of children (e.g., Hetherington & Ross, 1967) and adults (e.g., Aiken, 1968; Beeson, 1973; Gaynor, 1981). Surprisingly, a 24-hr delay of feedback has been reported to have a positive influence on learning, an outcome known as the delayed reinforcement effect (DRE) (e.g., Brackbill, Bravos, & Starr, 1962; Kulhavy & Anderson, 1972; Surber & Anderson, 1975). The mechanisms underlying the DRE appear to be related to the general beneficial effects of feedback, such as the correction of previously inaccurate assumptions and the reduction of inaccurate perseverative responding. The typical multiple-choice test may be an effective and practical assessment tool but it does not convert mistakes into new learning. Indeed, without corrective feedback, the learner likely exits an examination assuming that an incorrect response was actually correct; thus, an examination that does not employ feedback may promote misconceptions. A more optimal multiple-choice testing format would not only assess the learner's current level of understanding, but would also correct misunderstandings. That is, the test would teach as well as assess.

In a recent report, we described the benefits of an answer-until-

correct (AUC) multiple-choice procedure that provided immediate feedback and enabled, at instructor discretion, the assignment of partial credit for proximate knowledge—the Immediate Feedback Assessment Technique (IF AT) (Epstein, Epstein, & Brosvic, 2001). Performance on the IF AT was compared with performance on identical tests when answers were recorded on Scantron forms which provided neither feedback nor the opportunity to answer until correct. Participants used either IF AT or Scantron forms to respond to unit tests, and then all participants used only the Scantron form to respond to the final examination which contained some questions repeated from the earlier unit tests. Test scores on the unit tests did not differ between the two test formats because the learning that the IF AT promotes should be reflected in the cumulative correction of initially incorrect responses on the unit test items repeated on the final examination. As expected, participants tested with the IF AT on the unit tests correctly answered more of the final examination questions that had been repeated from earlier unit tests than did participants tested with Scantron forms. Similarly, participants tested with the IF AT correctly answered more of the final examination questions that they had previously answered incorrectly on the unit tests than did participants tested with Scantron forms. Approximately 60% of the errors initially made on unit tests when the IF AT was used were converted to correct answers on the final examination, whereas approximately 70% of the errors initially made on unit tests when Scantron forms were used were repeated on the final examination. These results were especially noteworthy because the feedback was immediate but the delay until the items were presented on the final examination ranged between 3 and 10 weeks.

The robustness of the IF AT as a means by which to correct previously inaccurate assumptions was replicated in additional studies conducted in our laboratory and prompted the studies described below. In comparison to our earlier reports, the testing situation in Studies 1 and 2 did not involve classroom assessment and test-retest delays were standardized at either 1 day or 1 week. These procedures permitted the comparison of performance on the IF AT and the Scantron forms when concerns over participant motivation and course grades were removed. In Study 1, the initial test and retest items were identical whereas in Study 2 the questions and answer options on the retest were conceptually similar but not identical to items used on the initial test. In each of these two studies, all participants completed the retest using only Scantron forms.

Study 3 was prompted by the results of pilot studies in which a computerized testing system that provided the benefits of the IF AT method was neither preferred by participants nor found to enhance retention. The implementation of internet-based testing is increasing, and although electronic testing procedures may provide a cost-effective and labor-reducing method of assessment, they currently do not provide feedback. Thus, in Study 3 an immediate feedback and answer-until-correct procedure was provided to all participants, half with the IF AT and half with the computerized testing system for the initial tests, whereas on

the final test all participants used Scantron forms. Despite considerable symmetry in visual and tactile input between the IF AT and the input device (mouse), keyboard, and screen, we hypothesized that the IF AT promotes a more active discovery process than that afforded by the clicking of an input device and is also more analogous to the traditional and contemporary classroom testing environments. Accordingly, we predicted that participants evaluated with the IF AT would demonstrate enhanced retention.

Study 1

Method

Participants. Fifty female and 20 male undergraduate participants enrolled in Introduction to Psychology courses served as voluntary participants and received extra credit for participation. The modal participant was a female liberal arts major, Caucasian, and in the first or second year of study.

Materials. The testing formats were identical to those described previously by Epstein et al. (2001). Briefly, the IF AT form is a multiple-choice answer form with rows and columns of rectangular answer spaces corresponding to the number of the examination questions and the answer options, respectively. Participants scraped off an opaque, waxy coating covering each option to indicate an answer selection. A star indicated a correct selection; a blank space indicated an incorrect answer. The placement of the star was randomized across questions. The Scantron form had the same number of rows and columns of blank answer spaces; a participant indicated an answer by darkening the appropriate space with a pencil. Both answer forms were commercially designed and commercially printed. The IF AT was prepared in eight versions so that the placement of the star could be varied, and a representative sample of the IF AT is presented in Figure 1.

Design and procedure. Participants completed a 20-item multiple-choice trivia test in small groups of 5 or fewer participants who were instructed to read each question, evaluate the response options, and select the correct answer. Thirty-three participants were randomly assigned to record their answers using Scantron forms. Thirty-seven participants were randomly assigned to record their answers using the IF AT form. The latter participants were informed that they would uncover a star if they were correct. In the event that their responses were not correct, they were instructed to reconsider questions and remaining response options and to continue responding until they made correct selections. Once all of the initial (Time 1) ratings were completed, one half of the participants in each group was randomly assigned to be retested after a delay of either 1 day or 1 week (Time 2). At each of these delay intervals, the same multiple-choice test questions were administered although the ordering of the questions and response options was altered to reduce response biases, and they were completed by all participants

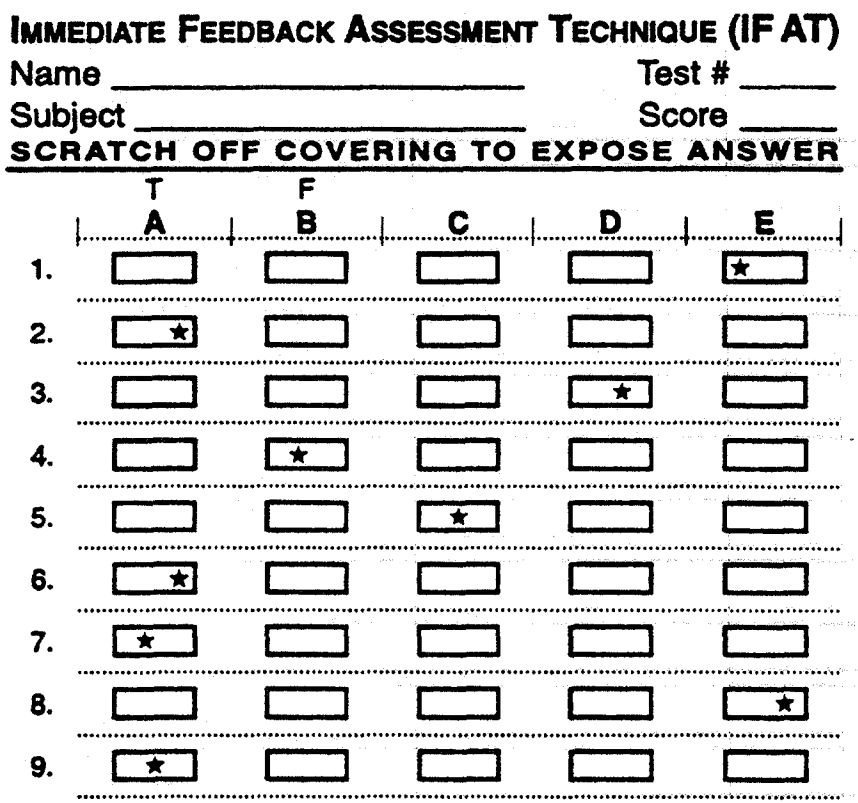

Figure 1. Sample portion of the Immediate Feedback Assessment Technique (IF AT) form. Trademark and patent are held by the senior author.

using Scantron forms. Thus, test format (IF AT, Scantron) served as the between-subjects factor whereas repeated testing (Time 1, Time 2) and delay (1 day, 1 week) served as the within-subjects factor. Although the IF AT method enables the assignment of partial credit (i.e., correct responding on the first attempt is assigned 100% of item credit whereas responding on the second, third, or fourth attempt could be assigned reduced percentages according to instructor discretion), this procedure was not used and the results described below were based upon the accuracy of initial responses.

Results

The mean number of correct responses for the IF AT and Scantron forms is presented in Figure 2 as a function of time of testing. The main and interaction effects were significant [all $F(1, 61) > 8.45$, all $p < .05$]. Scheffé comparisons indicated that mean scores at the initial test did not differ between the IF AT and Scantron forms. However, mean scores at the 1-day and 1-week delays were significantly higher for participants

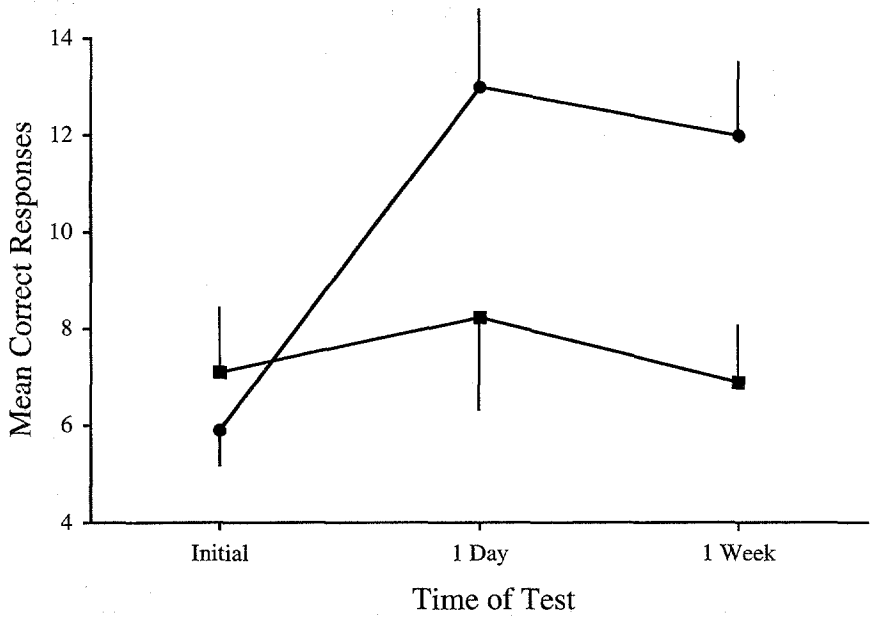

Figure 2. Mean correct responses for the IF AT (closed circles) and the Scantron (closed squares) groups for the initial test and the 1-day- and 1-week-delayed retention tests in Study 1.

evaluated with the IF AT than for participants evaluated with Scantron forms, the 1-day and 1-week scores were significantly higher than initial scores for participants evaluated with the IF AT, and the 1-day and 1-week scores did not differ from the initial scores for participants evaluated with Scantron forms (Scheffé comparisons, all $p < .05$).

The significantly higher mean scores at the 1-day and 1-week retests for the IF AT group were not related to between-group differences in initial scores; rather, they were related to the feedback that the IF AT method

Table 1

Conditional Probability (in percentages) of Test 2 Outcomes Given Test 1 Outcomes By Test Method and Delay Interval in Study 1

	Scantron		IF AT	
	Day	Week	Day	Week
Outcome Conditions				
Correct Time 2 / Correct Time 1	71.09	69.69	84.89	82.39
Correct Time 2 / Incorrect Time 1	14.62	11.53	57.79	48.44
Incorrect Time 2 / Correct Time 1	18.91	30.31	15.11	17.61
Incorrect Time 2 / Incorrect Time 1	85.38	88.47	42.21	51.56

provides. This conclusion is supported by the conditional probabilities of correct responding at the 1-day and 1-week delays, as seen in Table 1. The probabilities represent the four potential conditions generated by correct responding on the initial test (Time 1) and at the appropriate delay (Time 2 scores). The main and interaction effects for an analysis of variance similar to that described above were again significant [all $F(1, 61) > 11.17$, all $p < .05$]. Scheffé comparisons indicated no significant difference between the IF AT and Scantron groups in either the probability of correct responses for Time 2 questions that had been answered correctly at Time 1 or the probability of incorrect responses on Time 2 questions that had been answered correctly at Time 1 (all $p > .05$). However, there were significant differences in performance on Time 2 items as a function of test format (IF AT versus Scantron) and initial performance on test item (correct versus incorrect). Scheffé comparisons indicated that (a) participants evaluated with the IF AT correctly answered significantly more Time 2 questions that had initially been answered incorrectly at Time 1 than did participants evaluated with Scantron forms and that (b) participants evaluated with Scantron forms, at both delays, incorrectly answered significantly more Time 2 questions that had initially been answered incorrectly at Test 1 than did participants evaluated with IF AT forms (all $p < .05$).

Study 2

Method

Participants. Forty female and 20 male undergraduate participants enrolled in Introduction to Psychology courses served as voluntary participants and received extra credit for participation. As in Study 1 the modal participant was a female liberal arts major, Caucasian, and in the first or second year of study.

Materials. The IF AT and Scantron testing methods were identical to those described above in Study 1.

Design and procedure. Participants were given as much time as required to read a three-page article concerning extrasensory perception, and upon completion, to complete a 15-item multiple-choice test about information presented in the article. Two versions of this test with comparable wording were constructed. One half of the participants were randomly assigned to complete one of the two versions as their initial test (Time 1) using the IF AT form, whereas the other half completed one of the two versions using the Scantron form. Within each test group, one half of the participants were randomly assigned to return either 1 day or 1 week later (Time 2). At Time 2, all participants completed the version not initially taken and did so either 1 day or 1 week later using the Scantron form. The scoring and analysis procedures were identical to those described in Study 1.

Results

Performance on the initial and delay tests, both for the IF AT and for the Scantron groups did not differ as a function of the version of item wording [all $t < 0.67$, all $p > .05$]. Thus, the use of conceptually similar but differently worded tests items does not account for the performance differences described below.

The mean number of correct responses for the IF AT and Scantron tests is presented in Figure 3 as a function of time of testing. Test format

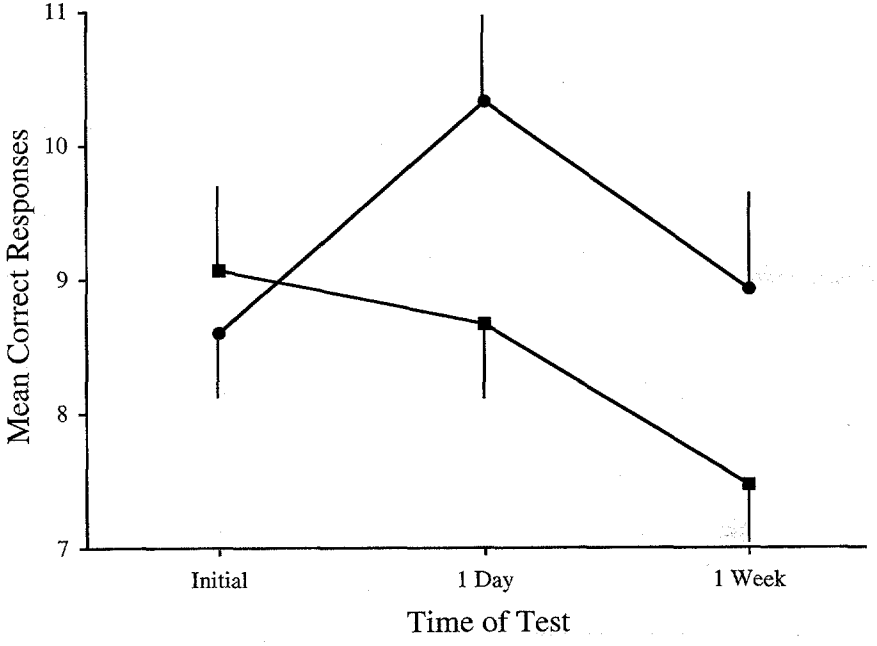

Figure 3. Mean correct responses for the IF AT (closed circles) and the Scantron (closed squares) groups for the initial test and the 1-day- and 1-week-delayed retention tests in Study 2.

(IF AT, Scantron) served as the between-subjects factors whereas repeated testing (Time 1, Time 2) and delay (1 day, 1 week) served as the within-subjects factors. The main and interaction effects were significant [all $F(1, 51) > 26.46$, all $p < .05$]. Scheffé comparisons indicated that mean scores at the initial test did not differ between the two test formats, that mean scores at the 1-day and 1-week delays were significantly higher for participants evaluated with IF AT forms, that 1-day scores were significantly higher than initial scores for participants evaluated with IF AT forms, and that 1-week scores were significantly lower than initial scores for participants evaluated with Scantron forms (all $p < .05$).

The percentage of change in correct responding from Time 1 to Time 2 is presented in Table 2 as a function of test format and length of delay. The main and interaction effects for an analysis of variance similar to that

Table 2

Mean Percentage of Change in Number of Correct Responses from Test 1 to Test 2 for Test Method and Delay Interval in Study 2

	Delay Interval	
	Day	Week
Test Method		
IF AT		
M	20.80	11.47
SD	23.45	25.68
Scantron		
M	5.33	-21.93
SD	18.14	17.55

described above were significant [all $F(1, 51) > 27.49$, all $p < .05$]. Scheffé comparisons for participants evaluated with the IF AT indicated that the percentage of change in correct responding was significantly greater at the 1-day and 1-week delays, and greater for the 1-day than for the 1-week delay (all $p < .05$). Scheffé comparisons for those evaluated with Scantron forms indicated that the percentage of change in correct responding was significantly lower for the 1-week than for the 1-day delay ($p < .05$). Thus, participants evaluated with Scantron forms demonstrated progressive declines in performance as a function of delay, especially after a 1-week delay, whereas participants evaluated with the IF AT demonstrated enhanced performance at both delay periods, with the greatest amount of improvement observed after a delay of 1 day (all $p < .05$).

Study 3

Method

Participants. Thirty female and 16 male undergraduate participants enrolled in Introduction to Psychology courses served as voluntary participants and received extra credit for participation. As in Studies 1 and 2 the modal participant was a female liberal arts major, Caucasian, and in the first or second year of study.

Materials. The IF AT testing was identical to that described above in Studies 1 and 2. The software program for the Macintosh PowerPC was written in the basic programming language.

Procedure and design. Participants were provided with as much time as needed to read a two-page article concerning obsessive-compulsive disorders, and upon completion, to complete a 14-item multiple-choice test about information presented in the article (Time 1). One half of the participants were randomly assigned to complete the test using IF AT forms whereas the others responded on a Macintosh PowerPC computer using a software program that provided immediate feedback for each answer option and permitted participants to continue answering until the correct answer was selected. One half of the participants in each group was randomly assigned to return either 1 day or 1 week later (Time 2),

and at that time, all participants used Scantron forms. The questions were identical to those taken initially although the order of the items and the answers were altered to reduce response bias.

Results

The mean number of correct responses for the IF AT and computerized tests is presented in Figure 4 as a function of time of

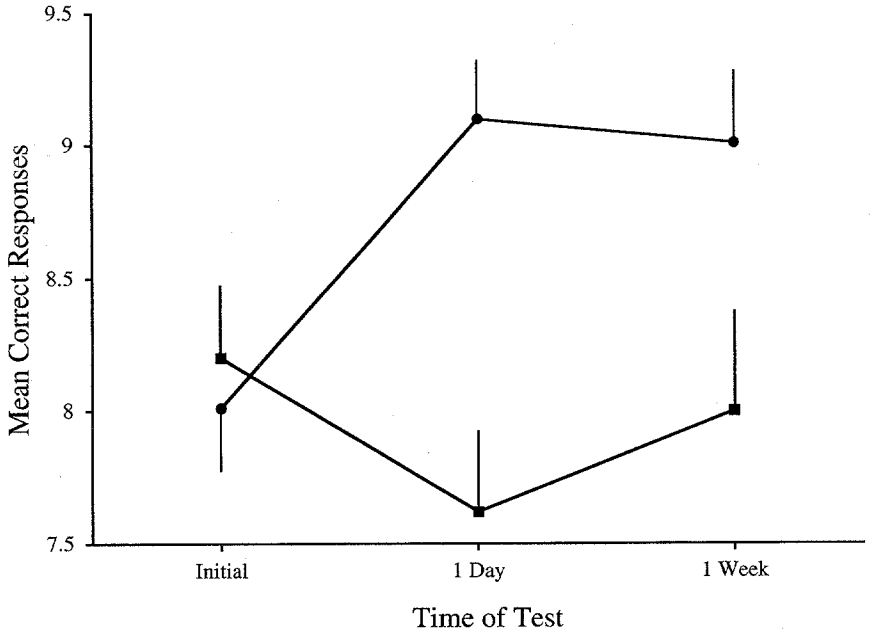

Figure 4. Mean correct responses for the IF AT (closed circles) and the computer (closed squares) groups for the initial test and the 1-day- and 1-week-delayed retention tests in Study 3.

testing. Test format (IF AT, Computer) served as the between-subjects factor whereas repeated testing (Time 1, Time 2) and delay (1 day, 1 week) served as the within-subjects factor. The main and interaction effects were significant [all $F(1, 40) > 7.03$, all $p < .05$]. Scheffé comparisons indicated that mean scores at the initial test did not differ between the two test formats, that mean scores at the 1-day and 1-week delays were significantly higher for participants evaluated with IF AT forms than for participants evaluated by computer, that 1-day and 1-week scores were significantly higher than initial scores for the IF AT participants, and that initial scores did not differ from scores at the 1-day- and 1-week-delayed test scores for participants evaluated by computer (all $p < .05$).

The percentage of change in correct responding, calculated by

Table 3

Mean Percentage of Change in Number of Correct Responses from Test 1 to Test 2 By Test Method Answer and Delay Interval in Study 3

Delay Interval	Day	Week
Test Method		
IF AT		
M	14.40	12.57
SD	25.46	30.12
Computer		
M	-7.40	-2.18
SD	16.23	8.44

comparing performance between Time 1 and Time 2, is shown in Table 3. The main and interaction effects were significant [all $F(1, 40) > 10.95$, all $p < .05$]. Scheffé comparisons for participants evaluated with the IF AT indicated that the percentage of change in correct responding on Time 2 did not differ between the two delay periods and that the percentage of change was significantly higher at the 1-day and the 1-week intervals than that observed at initial testing (all $p < .05$). Scheffé comparisons for participants evaluated with the computer indicated no significant differences in percentage of change between the initial and delayed tests.

Discussion

In Studies 1 and 2 mean test scores on initial tests did not differ between participants evaluated with the IF AT and participants evaluated with Scantron forms. This outcome was predicted because test scores were based upon initial responses; thus, the beneficial effects of feedback should not emerge until the delayed tests. On those tests, conducted after delays of 1 day or 1 week, mean test scores were significantly greater for participants evaluated with the IF AT. These increases were related to the feedback that the IF AT provides, as evidenced by the conditional probabilities of correct responding. Participants evaluated with the IF AT were able to correct initially inaccurate answer strategies, and during the delayed tests they accessed correct information to respond accurately to many of the unit test items that they had initially answered incorrectly. Participants evaluated on identical items with Scantron forms responded in the absence of corrective feedback, and at both delay intervals they continued to respond incorrectly. As discussed below, the percent increases for the IF AT group in Study 3 support the efficacy of the IF AT over a computer-based system providing identical levels of feedback. There may be factors that account for these differences, although they have yet to be identified.

The results of Studies 1 and 2 provide important replications and extensions of our previous study (Epstein et al., 2001). In that study, participants were evaluated on unit tests during the semester using either

IF AT or Scantron forms; the final test was completed by both groups using Scantron forms. Mean scores on each unit test did not differ between participants evaluated with either the IF AT or with Scantron forms, an outcome similar to those described above for Time 1 ratings in Studies 1 and 2. However, participants evaluated with IF AT forms on the unit tests correctly answered more of the final examination questions that had been repeated, especially when they had answered the same questions incorrectly on the unit tests, than did participants evaluated with Scantron forms. This latter outcome is similar to the results of conditional probabilities analyses reported for Studies 1 and 2. The convergence of the results is noteworthy as there were substantial differences in delay intervals, stimulus materials (actual classroom testing versus small group tests on nonclassroom materials), and motivational factors related to test outcomes. The IF AT method enhanced the retention of repeated items, especially those items that were not initially answered correctly. This outcome represents learning during the testing process through the correction of initially inaccurate assumptions—an outcome obtained with neither the Scantron form in Studies 1 and 2 nor computer-based testing in Study 3. The robustness of the IF AT method to correct inaccurate strategies for answering and the retention of this information over periods ranging from 1 day to 10 weeks suggest potential beneficial effects during preparation for graduate school subject matter entrance examinations and professional licensing examinations for which practice tests are available. The results of Study 3, however, have implications for how feedback should be provided.

In Study 3, participants reviewed materials similar to the materials used in Studies 1 and 2, responded using either the IF AT or a computer keyboard, and were tested again after delays of 1 day or 1 week. Unlike Studies 1 and 2, all participants received immediate feedback after each response and continued to respond until correct. The delivery of feedback by computer did not promote retention. One factor which has been shown to affect acquisition and retention and to be operative during computerized tests is participant involvement. Clariana, Ross, and Morris (1992) reported that the coupling of computerized multiple-choice testing with an answer-until-correct procedure produced the highest self-reports of active involvement and the most thorough processing of stimulus materials. The IF AT format also appears to engage participants in an active discovery process in which they are actually "discovering by uncovering" the answers and, at least in Study 3, computerized testing did not generate the same benefits. This conclusion is supported by posttest debriefings during which participants uniformly indicated preference for the IF AT, citing the importance of the active discovery process that it promotes and its similarity to normative classroom evaluations.

One issue yet to be resolved is the mechanism(s) by which feedback and the timing of its delivery facilitates the learning process. Rankin and Tepper (1978) reported that a 15-s delay promoted retention whereas Gaynor (1981) reported that delayed feedback reduced retention. Webb,

Stock, and McCarthy (1994) examined the effects of immediate and delayed feedback on the acquisition of general information multiple-choice items presented via computer. Each item stem was presented without response options, and participants rated confidence in their ability to answer prior to and after responding. Participants returned 1 and 6 days after initial testing, and feedback was provided, either immediately after responding or 24 hr later. Surprisingly, posttest scores were generally higher for participants receiving feedback after the 24-hr delay, and this outcome has become known as the "delayed reinforcement effect" or DRE. This outcome was replicated in a follow-up study in which delayed feedback increased test item study time and the conditional probability of correctly answering items on the posttest that had been incorrectly answered on the initial test. Peeck, van den Bosch, and Kreupeling (1985) reported nominal differences in performance between participants provided with immediate feedback and participants provided with delayed feedback, an outcome that provided minimal support for the DRE. Inspection of the Peeck et al. (1985) conditional probabilities of responding, calculated in the same manner as in the present studies, also provides no support for the DRE and its interference perseveration hypothesis, as the retention of initially incorrect responses did not preclude the acquisition of correct answers. These patterns of responding suggest that an awareness of initially inaccurate responses assisted with, rather than detracted from, the acquisition of correct responses when such responses were presented.

The results of the present studies indicate that multiple-choice tests which actively involve participants in the discovery of correct answers and provide immediate informative feedback in an answer-until-correct format promote acquisition and the retention of test materials. The IF AT format does all of the above; computerized testing does not do it as well; the Scantron form does not do it at all. Educators need to reconsider the utility of Scantron-type answer forms and computerized testing for assessing participant knowledge. Whereas multiple choice tests allow faculty to assess participant performance in large enrollment classes and to return examination results expeditiously, the testing format does not support new learning; in fact, Scantron-like answer forms appear to reinforce incorrect assumptions. The IF AT can be used with classes of any size; independent of class size, it retains the benefits of being an engaging medium that supports learning by providing reinforcing feedback for correct responses and corrective feedback for incorrect responses while involving the participant in a discovery process.

Skinner (1983) presented participants with materials about answer changing at the beginning of the semester and then examined the incidence of answer changes on a later assessment. In the absence of feedback, participants' initially inaccurate answers were more likely to be changed to correct answers, especially when there was high confidence in the decision to change. This outcome was questioned by Ramsey, Ramsey, and Barnes (1987) who not only reported the greatest gains for

answer changing on items of low difficulty when confidence for changing was high but also for items of varying difficulty when confidence was low. Skinner (1983) also reported that the tendency to change initial answers was more likely for female participants although causal factors could not be defined. Ramsey et al. (1987) reported that sex was unrelated to answer changing, even when coupled to ability, and this observation is consistent with studies conducted in our laboratory in which sex differences have not been observed. The incidence of answer changing with the IF AT was higher than that reported previously (e.g., Benjamin, Cavell, & Shallenberger, 1984), an inherent outcome of the answer-until-correct procedure. The correction of these initially inaccurate responses represents learning during the testing process and suggests that test takers refine decision making during testing. This learning process is analogous to the level of processing effect examined in the Lhyle and Kulhavy (1987) study in which participants rearranged the words within a feedback sentence in order to maximize its application to a learning situation. The discovery process that learners experienced with the IF AT and reported during posttest debriefings was one that required additional attention, concentration, and processing — factors shown to reduce errors (e.g., Benton, Glover, & Bruning, 1983; Glover, Bruning, & Plake, 1982).

It is generally agreed that the best tests are those that teach while assessing. We have demonstrated a powerful tool that allows instructors to assess sensitively while maximizing the probability that participants not only exit each item with the correct answer but also with information that transfers to later testing situations. In our past studies, the stimulus materials and tests were taken from regular classroom activities, randomly selected unit test items were repeated on the final exam, retest delays ranged between 3 to 10 weeks, and performance on the tests was consequated by the assignment of course grades. In our present studies, nonclassroom materials were used, delays were standardized at either 1 day or 1 week, and participant motivations included both interest and extra credit. Despite these noteworthy differences, the outcomes attributable to the IF AT method were consistent: greater retention and the correction of initially inaccurate answers to items repeated from prior exams. As seen in Study 2, this latter outcome was robust even when the repeated items were conceptually similar, albeit worded differently. Unlike the typical multiple-choice answer form, the IF AT does not foster the acquisition of incorrect information, and unlike computerized testing programs, the IF AT appears more directly to involve the participant in active information processing.

References

AIKEN, E. G. (1968). Delayed feedback effects on learning and retention of Morse Code symbols. *Psychological Reports*, 23, 723-730.

BEESON, R. O. (1973). Immediate knowledge of results and test performance. *Journal of Educational Research*, 66, 224-226.

BENJAMIN, L. T., CAVELL, T. A., & SHALLENBERGER, W. R. (1984). Staying with initial answers on objective tests: Is it a myth? *Teaching of Psychology*, 11, 133-141.

BENTON, S. L., GLOVER, J. A., & BRUNING, R. H. (1983). Levels of processing: Effects of numbers of decision on prose recall. *Journal of Educational Psychology*, 75, 382-390.

BRACKBILL, Y., BRAVOS, A., & STARR, R. H. (1962). Delay-improved retention of a difficult task. *Physiological Psychology*, 55, 947-952.

CLARIANA, R. B., ROSS, S. M., & MORRIS, G. R. (1992). The effects of different strategies using computer-administered multiple-choice questions as instruction. *Educational Technology Research & Development*, 39, 156-169.

EPSTEIN, M. L., EPSTEIN, B. B., & BROSVIC, G. M. (2001). Immediate feedback during academic testing. *Psychological Reports,* 88, 889-894

GAYNOR, P. (1981). The effect of feedback delay on retention of computer-based mathematical material. *Journal of Computer-Based Instruction*, 8, 28-34.

GLOVER, J. A., BRUNING, R. H., & PLAKE, B. S. (1982). Distinctiveness of encoding and recall of text materials. *Journal of Educational Psychology*, 14, 522-534.

HETHERINGTON, E. M., & ROSS, L. E. (1967). Discrimination learning by normal and retarded children under delay of reward and interpolated task conditions. *Child Development*, 38, 639-647.

KLUGER, A., & DENISI, A. (1998). Feedback interventions: Toward the understanding of a double-edged sword. *Current Directions in Psychological Science*, 7, 67-72.

KULHAVY, R. W., & ANDERSON, R. C. (1972). Delay-retention effect with multiple-choice tests. *Journal of Experimental Psychology*, 63, 505-512.

LHYLE, K. G., & KULHAVY, R. W. (1987). Feedback processing and error correction. *Journal of Educational Psychology*, 79, 320-322.

MISLEVY, R. J. (1991). A framework for studying differences between multiple-choice and free-response test items. Cited from R. D. Bennett & W. C. Ward (1993), *Construction vs. choice in cognitive reassessment.* Hillsdale, NJ: Erlbaum.

PEECK, J., VAN DEN BOSCH, A. B., & KREUPELING, W. J. (1985). Effects of informative feedback in relation to retention of initial responses. *Contemporary Educational Psychology*, 10, 303-315.

RAMSEY, P. H., RAMSEY, P. P., & BARNES, M. J. (1987). Effects of participant confidence and item difficulty on test score gains due to answer changing. *Teaching of Psychology*, 14, 206-210.

RANKIN, R. J., & TEPPER, T. (1978). Retention and delay of feedback in a computer assisted instruction task. *Journal of Experimental Education*, 46, 67-70.

SKINNER, N. F. (1983). Switching answers on multiple-choice questions: Shrewdness or shibboleth? *Teaching of Psychology*, 10, 220-222.

SURBER, J. R., & ANDERSON, R. C. (1975). Delay-retention effect in natural classroom settings. *Journal of Educational Psychology*, 7, 170-173.

WEBB, J. M., STOCK, W. A., & MCCARTHY, M. T. (1994). The effects of feedback timing on learning facts: The role of response confidence. *Contemporary Educational Psychology*, 19, 251-265.