

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. Mitlevy (1991) discusses the origins and explosive growth in multiple-choice testing since World War I. 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

IMMEDIATE FEEDBACK ASSESSMENT TECHNIQUE (IF AT)

Name _____

Test # _____

Subject _____

Score _____

SCRATCH OFF COVERING TO EXPOSE ANSWER

	T	F	C	D	E
	A	B	C	D	E
1.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
2.	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
3.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
4.	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
5.	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
6.	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
7.	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
8.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
9.	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

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

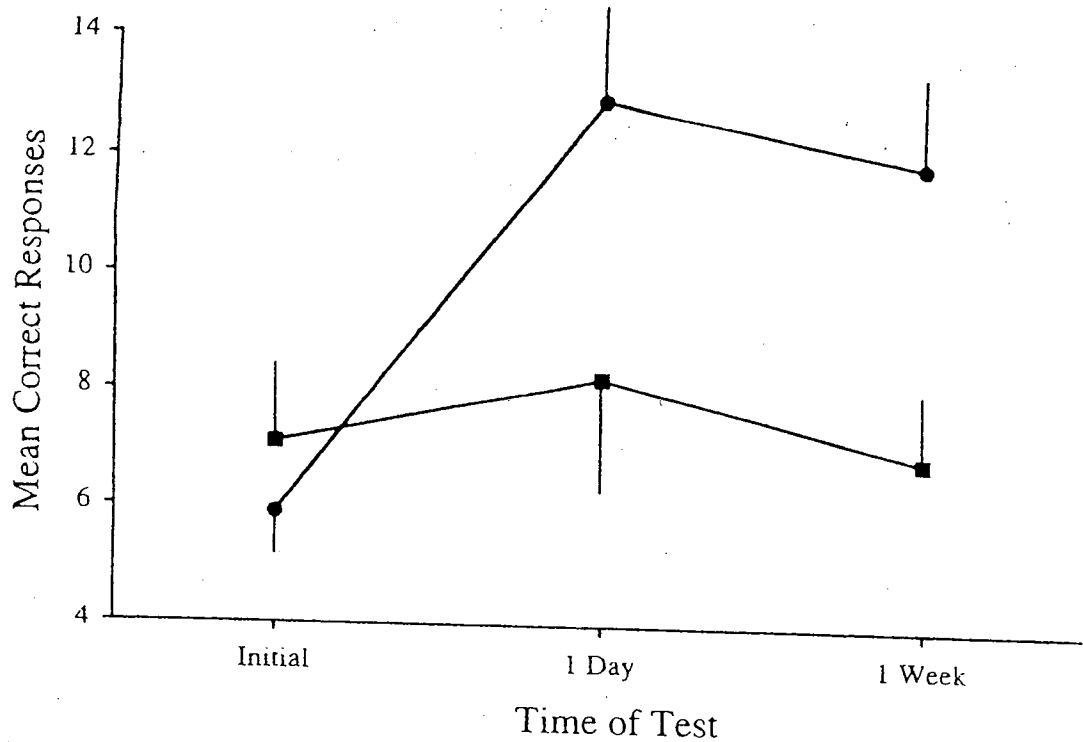


Figure 2. Mean correct responses for the IF AT (closed circles) and the Scantron (closed squares) groups for the initial test and the 1-day- and 1-week-delayed retention tests in Study 1.

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

Outcome Conditions	Scantron		IF AT	
	Day	Week	Day	Week
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

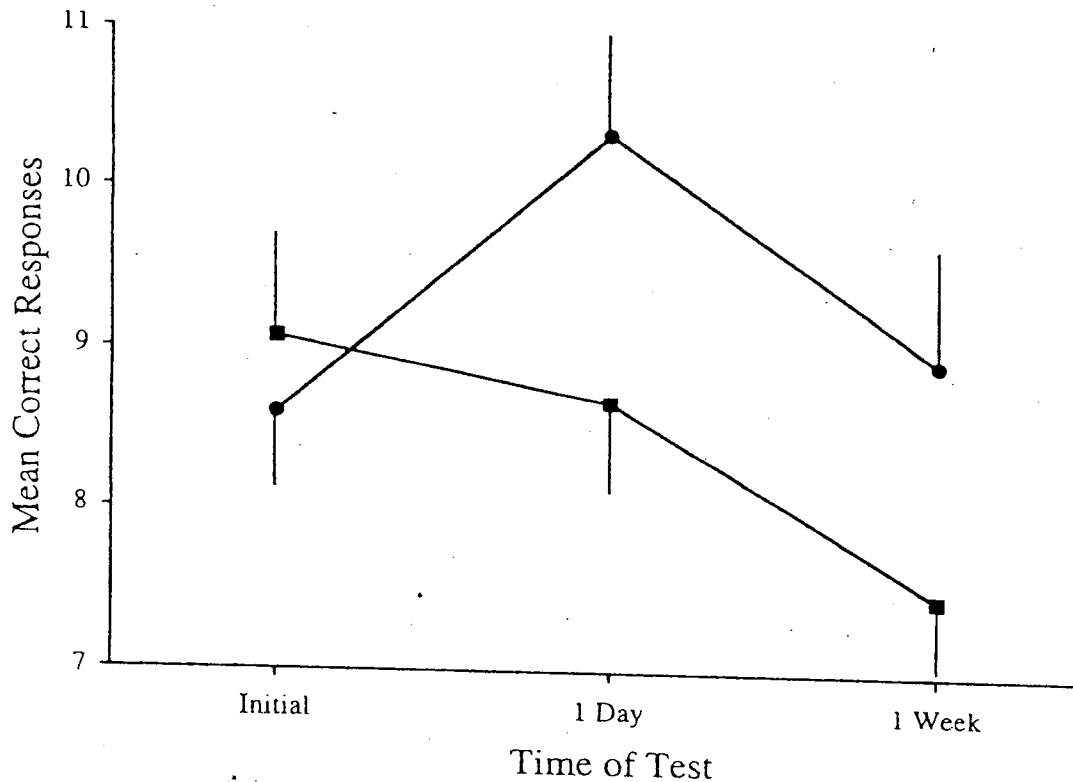


Figure 3. Mean correct responses for the IF AT (closed circles) and the Scantron (closed squares) groups for the initial test and the 1-day- and 1-week-delayed retention tests in Study 2.

(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

