Provision of Feedback During Preparation For Academic Testing: Learning is Enhanced by Immediate But Not Delayed Feedback

Roberta E. Dihoff (Rowan University), Gary M. Brosvic (Rider University), and Michael L. Epstein (Rider University)

Address correspondence to Gary M. Brosvic, Department of Psychology, Rider University, Lawrenceville NJ 08648. Email to Brosvic@Rider.edu.

Abstract

Students prepared for classroom examinations by completing practice tests, with selected items from these practice tests repeated, in either the original or in a modified wording, on classroom examinations and a final examination. The availability of immediate self-corrective feedback on Study 1 practice tests (0, 3, or 6 practice tests) was varied, while in Study 2, the timing of feedback provided during practice tests (immediate, end of test, 24-hour delay, control) was varied. Performance on examinations was elevated by the provision of immediate feedback on practice tests in both studies, especially when test items were presented in their original wording, with some generalization observed on items presented in a modified wording. Predictions made in accordance with the interference-perseveration hypothesis and the delay-retention effect were not supported. These results demonstrate considerable potential for immediate self-corrective feedback, delivered during test preparation through the Immediate Feedback Assessment Technique, to enhance performance on classroom examinations and to promote the retention of factual information during the academic semester.

Provision of Feedback During Preparation For Academic Testing: Learning is Enhanced by Immediate But Not Delayed Feedback

During the past five years Epstein and his colleagues have been refining and validating an assessment procedure known as the Immediate Feedback Assessment Technique, or IF AT (Dihoff, Brosvic, & Epstein, 2003; Epstein, Epstein, & Brosvic, 2001; Epstein, Lazarus, Calvano, Matthews, Hendel, Epstein, & Brosvic, 2002; Epstein, Brosvic, Dihoff, Lazarus, & Costner, 2003). The IF AT embodies the theoretical and practical foundations of the teaching-testing machines described by Pressey (1926) and Skinner (1958), transforming the passive receiver of information into the active demonstrator of skills and knowledge. The IF AT form (see Figure 1) is a multiple-choice answer sheet with rows of rectangular answer spaces (e.g., A, B, C, D) that is nearly identical in layout to the ubiquitous machine-scored answer sheet available from Scantron Corporation. Participants scrape off an opaque, waxy coating covering an answer space on the IF AT form to record their answer. If a symbol (e.g., a star) is printed beneath the covering the student receives instant feedback that a correct choice was made; the absence of a symbol provides instant feedback that an incorrect choice was made. However, rather than simply exiting the question, the student reviews the remaining response options, continues to respond until the correct answer is discovered (a self-correction procedure), and thus exits each question with the correct answer.

Although there is widespread agreement that learning is facilitated by feedback there is little agreement as to the type of feedback that is most effective (Robin, 1978). Proponents of the delay-retention effect (DRE), for example, report that the imposition of a 24-hr delay between the completion of a test and the provision and review of correct solutions is more likely to facilitate retention than

the provision of immediate feedback. In the early 1960's, Brackbill and her associates (Brackbill, Bravos, & Starr, 1962) demonstrated that delayed feedback across brief intervals promoted the retention of meaningful material. This outcome was also observed when feedback was delayed for 1 to 2 days and retention intervals were lengthened to 7 days (e.g., Kulhavy & Anderson, 1972; O'Neill, Rasor, & Bartz, 1976; Surber & Anderson, 1975). Proponents of delayed feedback typically also adhere to the interference-perseveration hypothesis proposed by Kulhavy and Anderson (1972): initial errors do not compete with to-be-learned correct responses if corrective information is delayed, because errors are likely to be forgotten, and thus they do not interfere with retention.

Corrective feedback on classroom examinations, in the absence of computers, cannot be provided until the examination has been completed, whereas the conditions and equipment within the laboratory permit the immediate delivery of corrective feedback on an item-by-item basis. The development of the IF AT now provides the practical means through which immediate feedback can be provided in the classroom without reliance upon elaborate technology, and it also permits the direct comparison of the effects of immediate and delayed feedback on classroom learning. For example, in Dihoff et al. (2003), students completed classroom examinations using response formats that provided feedback after each response, at the end of a test, or after a 24-hr delay. Students demonstrated the most recall, the most accurate identification of initial responses, the most confidence in their answers, and the least amount of perseverative incorrect responding on those final examination items that were originally responded to when immediate feedback was provided. These same students demonstrated less recall, less identification accuracy, less confidence in their answers, and more perseverative incorrect responding on

those final examination items that were originally responded to when either end of test or delayed feedback had been provided.

The studies conducted to date in our laboratories and classrooms have used the IF AT form as a tool to examine how the opportunity to answer until correct, coupled with immediate self-corrective feedback, enhances student learning. It is of interest, therefore, to examine the roles that feedback may play in the preparation of students for classroom examinations. Thus, in Study 1, we sought to examine the retention of learning when immediate self-corrective feedback was provided on 0, 3, or 6 practice tests, while in Study 2, we sought to examine how the timing of feedback (immediate, end of test, 24-hr delay) on practice tests influenced retention. In each study, selected items were carried over from practice tests to classroom examinations and to the final examination, using either the original or modified wording. We predicted that 1) the performance of students provided with feedback in both studies would be superior to that of no-feedback controls, especially when feedback was immediate, 2) higher levels of retention would be demonstrated by students completing classroom and final examination items worded identically to items on the practice tests, and 3) feedback would interact with item-wording, with the most retention demonstrated when immediate feedback was provided during practice tests and the wording of items was identical across the practice tests and the classroom and final examinations. The confirmation of these hypotheses, as described below, raise considerable implications for how practice tests, such as publisher-supplied study guides and web-based testing, could be structured to maximize learning and retention.

Study 1

Proactive Use of Feedback During Preparation for Academic Testing

Methods

Participants. Participants included 47 male and 73 female students enrolled in an undergraduate course. The modal participant was a caucasian female, majoring in the arts and sciences, with 60 or fewer credit hours completed at the time of the study.

Materials. Six out-of-class practice tests were prepared from a publisher-supplied test bank, each with 40 questions, and each question with four response options (i.e., A, B, C, D). Six classroom examinations were also prepared from a publisher-supplied test bank, each with 50 questions, and each question with four response options (i.e., A, B, C, D). Twenty items on each classroom examination were specifically related to items included in its respective practice test, with 10 items reproduced verbatim with 10 other items containing identical factual information that had been completely reworded. The final examination consisted of 120 items which included 5 of the verbatim and 5 of the reworded items used on each classroom examination plus 60 entirely new items. The reworded items were reviewed by a panel of students and faculty not involved in the present study, and were found to be of equivalent difficulty and linguistic complexity.

Design and Procedures

Immediate feedback was not available on any practice test for controls (n = 40), but in the experimental conditions, it was available on either 3 (n = 40) or 6 (n = 40) of the practice tests. Forty participants were randomly assigned to receive feedback on 3 practice tests in accordance with one of four orders of administration generated by the combination of IF AT (I) and Scantron (S) forms: I-I-I-S-S-S, S-S-

S-I-I-I, I-S-I-S-I-S, and S-I-S-I-S-I. Participants in the control condition recorded their responses to practice test items using Scantron forms, with these forms providing neither affirmation of correct responding and self-corrective information nor the opportunity to respond until correct. Participants in the immediate feedback conditions recorded their responses to practice tests using IF AT forms, with these forms providing self-corrective information, affirmation of correct responding, and the opportunity to respond until correct. Practice tests were provided 72 hours before the administration of in-class examinations, and each was returned to the experimenter before beginning the classroom examination. Participants reported the amount of study time prior to taking the classroom examination. The final examination was administered 2 weeks after the sixth classroom examination, with all participants recording their responses on Scantron forms. Once the final examination was completed, participants reviewed each final examination item and identified their initial responses, both those known to have been correct and those known to have been incorrect, and then rated confidence in the accuracy of their identifications on a 100-point scale ranging from 0 (no confidence) to 100 (complete confidence). Each participant then completed a 15-item questionnaire assessing ease of understanding and ease of completing response requirements, perceived fairness of and preference for an answer-until-correct procedure, and involvement in the test-taking process, as described previously by Epstein and Brosvic (2002). Participants receiving feedback on three of the six practice tests responded to these 15 items, separately, to evaluate both the immediate feedback and the Scantron procedures. Performance on the items carried over from practice tests to classroom examinations, and from classroom examinations to the final examination, served as the primary measure of recall. 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 correct responding on the second, third, or fourth attempt may 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

There were no differences in any dependent measure described below as a function of sex of participant [all $F < 1$, all $p > .5$]. There were no differences in either the amount of self-reported study time or practice test scores as a function of the number of practice tests on which feedback was provided [all $F < 1$, all $p > .5$]. This latter outcome was predicted as the beneficial effects of self-corrective information should not be observed until the classroom and final examinations. The analyses described below are based upon overall classroom and final examination scores that included items carried over from the practice tests, with a separate analysis of performance on just the items carried over from the practice tests also provided. No differences in the aggregate performance of participants receiving feedback on 3 of the 6 practice test were observed on either the classroom examinations or the final examination as a function of the order of tests on which feedback was provided [all $F < 1$, all $p > .5$]. Performance on the classroom examinations, within each of the four orders of administration, was significantly higher when feedback had been provided [all $F > 11.27$, all $p < .0001$]. It was reasonable, therefore, to dichotomize the performance of these participants into summary measures representing when feedback was available, and when feedback was not available, on the practice tests.

Classroom Examinations and Final Examination

Potential differences in performance on classroom and final examinations were examined using separate analyses of variance with the number of practice tests on which feedback was available as the between-subject factor. Significant differences were observed for scores on each classroom examination and the final examination (see Figure 2), and for the scores aggregated across classroom examinations [all $F > 19.02$, all $p < .0001$]. Scheffe comparisons indicated that scores on each classroom examination and the final examination were highest for participants receiving feedback [all $p < .001$]. Identical analyses were then completed on classroom and final examination scores after the items carried over from the practice tests were removed, with no performance differences observed in these analyses as a function of the number of practice tests on which feedback was provided [all $F < 1$, all $p > .5$]. This outcome was expected as feedback should not enhance performance on test items with which participants have not had prior experience.

Analysis of Retention and Item Wording

Potential differences in performance on the items carried over from the practice tests to the classroom examinations, and from the classroom examinations to the final examination, were examined using repeated-measures analyses of variance with the number of practice tests on which feedback was available as the between-subject factor and question wording and repeated administration as within-subject factors.

Percent correct responding on the items carried over with original wording and with modified wording is presented in Figures 3 and 4, respectively. The main effects of the number of practice tests on which feedback was available and question

wording, and their interaction, were significant [all $F > 26.83$, all $p < .0001$]. Scheffe comparison indicated percent correct responding was higher 1) for items with identical than for items with modified wording on both the classroom examinations and the final examination, 2) when immediate feedback was provided during a practice test, and 3) when immediate feedback was provided on practice tests and test items were presented on classroom examinations and the final examination in their original wording [all $p < .001$]. No differences in aggregate performance was observed on test items not carried over from the practice tests to the classroom and final examinations as a function of the number of practice tests on which feedback was provided [$F < 1$, $p > .5$]. Thus, the provision of feedback during each practice test appears to exert a greater influence when interacting with the processes of recognition and identification, with significantly less robust effects observed when interacting with the processes of discrimination and generalization to items with modified wording.

Conditional Probabilities

In the analyses described below, it became apparent that the outcomes observed for participants receiving feedback on three of the six practice tests needed to be dichotomized into summary measures that distinguished when immediate feedback was (FB - 3/3) and when immediate feedback was not available (FB - 0/3).

Differences in perseverative responding were evaluated by determining the conditional probabilities of correct responding on the first (practice test), second (classroom examination), and third (final examination) administrations of test items using an analysis of variance with the number of practice tests on which feedback was provided as the between-subject factor and item wording and repeated administration as within-subject factors. Significant differences in conditional

probabilities were observed for the main and interactive effects of the number of practice tests on which feedback was provided and item wording [all $F > 25.64$, all $p < .001$]. Conditional probabilities for items of identical and modified wording are presented in Tables 1 and 2, respectively. The results of Scheffe comparisons were consistent for test items, independent of their wording, and thus an integrated explanation is presented.

The likelihood of responding correctly on the initial and subsequent administrations of an item was higher when feedback was provided 1) during each practice test and 2) on 3 practice tests than on none [all $p < .001$]. The likelihood of responding correctly on subsequent administrations of an item, after having responded incorrectly on the initial administration, was higher when feedback was provided 1) during each practice test and 2) on 3 practice tests than on none [all $p < .001$]. The likelihood of responding incorrectly on subsequent administrations of an item, after having responded correctly on the initial administration, was lowest when feedback was provided 1) during each practice test and 2) lower when feedback was provided on 3 practice tests than on none [all $p < .001$]. The likelihood of responding incorrectly to the same item across repeated administrations was lowest when feedback was provided 1) during each practice test and 2) lower when feedback was provided on 3 practice tests than on none [all $p < .001$].

Across the three administrations consistent differences were observed as a function of item wording for participants receiving feedback on 3 or more practice tests. In comparison, no differences were observed when feedback was not available, an expected outcome as the lack of self-corrective information should not differentially affect performance. The comparison of conditional probability values presented in Tables 1 and 2 indicate that the provision of feedback on each practice

test, when item wording was identical, increased the likelihood of responding 1) correctly on the initial and subsequent administrations of an item and 2) correctly on subsequent administrations of an item after having responded incorrectly on the initial administration [all $p < .001$]. In contrast, the likelihood of responding incorrectly to the same item across repeated administrations was significantly higher when item wording was modified [$p < .001$]. The provision of feedback on 3 practice tests increased the likelihood of responding correctly on the initial and subsequent administrations of an item when test items were presented in their original wording [$p < .001$].

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 (Kulhavy & Anderson, 1972; Surber & Anderson, 1975). These two putative mechanisms were evaluated by a further analysis of responses to the items on the final examination: after completing the final examination each participant was requested to identify those final examination items that they recalled answering, either correctly or incorrectly, on the item's respective practice test, and to then report the degree of confidence in each identification. The percentage of items identified as being answered correctly and answered incorrectly on practice tests was examined using separate ANOVAs with the number of practice tests on which feedback was provided as the between-subject factor and item wording as the within-subject factor. Significant main and interactive effects were observed for the number of practice tests on which feedback was provided and the wording of test items in each analysis [all $F = 43.05$, all $p < .0001$]. Scheffe comparisons indicated that accuracy of identifying both initially-correct (see Figure 5) and initially-incorrect (see Figure 6)

responses was 1) highest when feedback was provided on each practice test, 2) higher when feedback was provided on 3 practice tests than on none, and 3) higher when immediate feedback was provided on practice tests and test items were presented in their original wording [all $\underline{p} < .001$].

Potential differences in confidence ratings for identifications (see Figure 7) were examined using an ANOVA with the number of practice tests on which feedback was provided as the between-subject factor and for item wording as the within-subject factor. Significant main and interactive effects were observed for the number of practice tests on which feedback was provided and item wording [all $\underline{F} > 23.02$, all $\underline{p} < .0001$]. Scheffe comparisons indicated that confidence levels were higher when 1) feedback was provided on each practice test, 2) feedback was provided on 3 practice tests rather than on none, and 3) immediate feedback was provided on practice tests and test items were presented in their original wording [all $\underline{p} < .001$]. Collectively, the results of these analyses indicate that the provision of immediate feedback enabled participants to recall more of their initial responses, both correct and incorrect, and to be more confident in the identification of their responses.

Participants' evaluations of their experimental condition were expressed on a brief questionnaire upon the conclusion of the final examination. The six scales described by Epstein and Brosvic (2002) were verified through factor analysis, with potential differences in scale scores examined using analyses of variance with the number of practice tests on which feedback was available as the between-subject factor. Mean responses on the scales (see Table 3) measuring test anxiety and time requirements did not differ as a function of the number of practice tests on which feedback was provided (all $\underline{F} < 1$, all $\underline{p} > .5$). Mean responses on the scales

measuring satisfaction with response format, clarity of response requirements, the desirability of the response form, and the benefits of testing differed significantly as a function of the number of practice tests on which feedback was available (all $F > 8.73$, all $p < .001$). Scheffe comparisons indicated that mean scores on these latter four scales were highest when feedback was provided on each practice test [all $p < .001$].

Study 2

Differential Effects of Immediate and Delayed Feedback During Preparation for Academic Testing

Methods

Participants. Participants included 32 male and 48 female students enrolled in an undergraduate level course. The modal participant was a caucasian female, majoring in the arts and sciences, with 60 or fewer credit hours completed at the time of the study.

Materials. Four out-of-class practice tests were prepared from a publisher-supplied test bank, each with 40 questions, and each question with four response options (i.e., A, B, C, D). Four classroom examinations were also prepared from a publisher-supplied test bank, each with 50 questions, and each question with four response options (i.e., A, B, C, D). Twenty items on each classroom examination were specifically related to items included from its respective practice test, with 10 items reproduced verbatim and 10 other items that contained identical factual information but had been completely reworded. The final examination consisted of 80 items which included 5 of the verbatim and 5 of the reworded items used on each classroom examination plus 40 entirely new items. The reworded items were reviewed by a panel of students and faculty not involved in the present study, and

were found to be of equivalent difficulty and linguistic complexity.

Design and Procedures. Twenty students were randomly-assigned to each of the experimental conditions described below. A latin squares design was used so that participants progressed through the response formats in a counterbalanced order. Participants reported the amount of study time prior to taking the classroom examination. Answers in the Control condition were recorded with a pencil on a Scantron form. Answers in the End of Test Feedback condition were also recorded with a pencil on a Scantron form. However, upon the completion of the test, all writing implements were removed and the participants were permitted to review the examination, the correct solutions, and their answer sheets for 30 minutes. Participants in the other conditions remained seated during this time and worked on non-course materials. Answers in the Delayed Feedback condition were also recorded with a pencil on a Scantron form. However, on the following day, these participants reviewed the examination, the correct solutions, and their corrected answer sheets for 30 minutes while the other participants remained seated and worked on non-course materials. Answers in the Immediate Feedback condition were recorded using the IF AT form (E3 Corporation) which enabled participants to receive immediate affirming or corrective feedback and to respond until the correct answer was discovered. The final examination was completed 2 weeks after completion of the fourth quiz, and at that time, all participants used Scantron forms to record their answers. Once the final examination was completed, participants reviewed the final examination items and identified their initial responses, both correct and incorrect, and then rated confidence in their decisions according to the procedures described in Study 1. Participants then completed the 15-item questionnaire described above in Study 1. 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 correct responding on the second, third, or fourth attempt may 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

There were no differences in any dependent measure described below as a function of sex of participant [all $F < 1$, all $p > .5$]. There were no differences in either the amount of self-reported study time or practice test scores as a function of the timing of feedback [all $F < 1$, all $p > .5$]. This latter outcome, consistent with the results of Study 1, was predicted as the beneficial effects of self-corrective information should not be observed until the classroom and final examinations. The analyses described below are based upon overall classroom examination scores that included items carried over from the practice tests. A separate presentation of performance on only those items carried over from the practice tests is also presented.

Classroom Examinations and Final Examination

Potential differences in scores on the classroom and final examinations were examined using a repeated-measures ANOVA with the timing of feedback as the within-subject factor. Significant differences were observed in scores on each examination (see Figure 8) as a function of the timing of feedback [$F = 35.87$, $p < .0001$]. Scheffe comparisons indicated scores on each classroom examination and the final examination were higher when 1) immediate feedback was provided, and 2) either delayed or end of test feedback rather than when a Scantron form was provided [all $p < .001$]. Identical analyses were completed on classroom and final

examination scores after the items carried over from the practice tests were removed, with no performance differences observed as a function of the timing of feedback [all $F < 1$, all $p > .5$].

Analysis of Retention and Item Wording

Potential differences in performance on the items carried over from the practice tests to the classroom examinations, and from the classroom examinations to the final examination, were examined using a repeated-measures analysis of variance with the timing of feedback, question wording, and repeated administration as within-subject factors.

Percent correct responding on the items carried over in either their original or modified wording is presented in Figures 9 and 10, respectively. The main effects of the timing of feedback and question wording, and their interaction, were significant [all $F > 7.63$, all $p < .001$]. Scheffe comparisons indicated that percent correct responding was higher 1) when immediate feedback was provided, 2) for items with identical than for items with modified wording on both the classroom examinations and the final examination, and 3) when either end of test or delayed feedback rather than a Scantron form was provided [all $p < .001$]. No differences in performance on those items not carried over from either a practice test to a classroom examination, or from a classroom examination to the final examination, were observed as a function of the timing of feedback, as these items had not received the varying treatments [all $F < 1$, all $p > .5$]. Collectively, these results suggest that the feedback provided by the IF AT exerts a greater influence when interacting with the processes of recognition and identification, with significantly less robust effects observed when interacting with the processes of discrimination and generalization while solving items with modified wording.

Conditional Probabilities

Differences in perseverative responding were evaluated by determining the conditional probabilities of correct responding on the first (practice test), second (classroom examination), and third (final examination) administrations of each item using a repeated-measures analysis of variance with the timing of feedback, question wording, and repeated administration as within-subject factors. Significant differences in conditional probabilities were observed as a function of the main and interactive effects of the timing of feedback and item wording [all $F > 34.72$, all $p < .001$]. Conditional probabilities for items with identical and with modified wording are presented in Tables 4 and 5, respectively, as a function of the timing of feedback. The results of Scheffe comparison were consistent, independent of item wording, and thus an integrated explanation is presented.

The likelihood of concurrence of correct responding on the initial and subsequent administrations of an item was 1) highest when immediate feedback was provided, and 2) higher when either delayed or end of test feedback rather than a Scantron form was provided [all $p < .001$]. The likelihood of responding correctly on subsequent administrations of an item, after having responded incorrectly on the initial administration, was 1) highest when immediate feedback was provided, and 2) higher when either delayed or end of test feedback rather than a Scantron form was provided [all $p < .001$]. The likelihood of responding incorrectly on subsequent administrations of an item, after having responded correctly on the initial administration, was 1) lowest when immediate feedback was provided, and 2) lower when either delayed or end of test feedback rather than a Scantron form was provided [all $p < .001$]. The likelihood of responding incorrectly to the same item across repeated administrations was 1) lowest when immediate feedback was

provided, and 2) lower when either delayed or end of test feedback rather than a Scantron form was provided [all $p < .001$]. Consistent differences were observed across the three administrations (practice test, classroom examination, final examination) as a function of item wording, especially when immediate feedback was provided; however, no differences were observed when Scantron forms were used, an expected outcome given the absence of self-corrective information. Within the IF AT condition, the likelihood of responding correctly on 1) the initial and subsequent administrations of an item, and 2) subsequent administrations of an item, after having responded incorrectly on the initial administration, was highest when test items were presented in their original wording [all $p < .0001$]. Within the End of Test Feedback and Delayed Feedback conditions, the likelihood of responding 1) correctly on the initial and subsequent administrations of an item was higher when the item wording was identical, and 2) incorrectly on subsequent administrations of an item, after having responded correctly on the initial administration, was significantly higher when the wording was modified [all $p < .001$].

As in Study 1, the mechanisms underlying the DRE were evaluated by requesting participants, after completing the final examination, to identify those final examination items that they recalled answering, correctly and incorrectly, on the practice test, and to then report the degree of confidence in their identifications. The percentage of items identified as being answered correctly (see Figure 11) and answered incorrectly (see Figure 12) on the initial examination were examined using separate ANOVAs with the timing of feedback and item wording as the within-subject factors, with significant main and interactive effects observed [all $F > 6.41$, all $p < .004$]. Scheffe comparisons indicated that identification accuracy was 1) highest when immediate feedback was provided, 2) higher when either delayed or

end of test feedback rather than a Scantron form was provided, and 3) higher within each feedback condition when the items were presented in their original wording [all $\underline{p} < .001$]. Potential differences in confidence ratings (see Figure 13) were examined using an ANOVA with the timing of feedback and item wording as within-subject factors. Significant main and interactive effects were observed for the timing of feedback and item wording [all $\underline{F} = 16.55$, all $\underline{p} < .0001$]. Scheffe comparisons indicated that confidence levels were 1) highest when immediate feedback was provided, 2) higher when either delayed or end of test feedback rather than a Scantron form was provided, and 3) higher for each feedback condition when test items were presented in their original wording [all $\underline{p} < .001$]. Collectively, the results of these analyses indicate that participants were most likely to correctly identify their initial responses, both correct and incorrect, and to be more confident in their identifications, when they had been provided with immediate self-corrective information.

Participants' evaluations of each experimental condition were expressed on a brief questionnaire that assessed overall perceptions of each response format upon the conclusion of the final examination. The six scales described previously in Study 1, after replication through factor analysis, were formed and descriptive statistics are presented in Table 6. Potential differences in scale scores were examined using analyses of variance with the timing of feedback as the within-subject factor. Mean responses on the scales measuring test anxiety and time requirements did not differ as a function of the timing of feedback (all $\underline{F} < 1$, all $\underline{p} > .5$). Mean responses on the scales measuring satisfaction with response format, clarity of response requirements, the desirability of the the timing of feedback, and the benefits of testing differed significantly as a function of timing of feedback (all $\underline{F} > 4.39$, all $\underline{p} < .001$), with

Scheffe comparisons indicating significantly higher scores when immediate feedback was provided [all $p < .001$]. Scheffe comparisons also indicated higher scores on the benefits of testing item when either delayed or end of test feedback rather than a Scantron form was provided [all $p < .001$].

Discussion

As stated earlier, teaching-testing machines introduced by Pressey (1926) and Skinner (1958) transformed the role of the student from a passive receiver of information to an active demonstrator of skills and knowledge. Programmed instruction by teaching machines was, in part, intended to maintain vigilance during the testing process, the benefits of which have been demonstrated across a number of tasks (see Kritch & Bostow, 1998; Miller & Malott, 1997; Tudor, 1995). The IF AT is an extension of these early teaching innovations, providing immediate self-corrective information and the opportunity to answer until correct, and thus engaging the learner in the teaching-testing process.

The robust effects observed when immediate feedback was provided during practice tests suggests an optimal window within which immediate, self-corrective information should be delivered. The results of Study 1 highlight the importance of providing immediate feedback while students prepare for formal classroom assessments, with the greatest enhancement in examination scores and the largest reductions in initially-inaccurate perseverative responding observed when feedback was available during each practice test. These outcomes were replicated in Study 2 in which participants were exposed to each feedback condition. The provision of immediate feedback, in both studies, promoted the most retention and the most accurate identification of initial responses, both correct and incorrect.

The combination of immediate feedback and the opportunity to answer until

correct has typically been available only in laboratory studies employing computer-assisted instruction, but now it is available through the IF AT for regular classroom examinations that include multiple-choice and alternative-choice questions. In one of our prior studies (Epstein et al., 2002) students receiving feedback from the IF AT demonstrated higher levels of retention than students completing identical items but receiving feedback from a computerized answer-until-correct procedure. These same students reported that the IF AT was a more engaging medium that was easy to complete and more natural to the testing process than the computer. The IF AT has noteworthy advantages in that it can be completed in any testing environment, that it requires no automation, and that it can be successfully completed by students with developmental disabilities possessing sufficient motor control to meet response requirements. Computer-assisted programming, alternatively, can be constructed to permit branching and the presentation of support and ancillary materials that a paper-and-pencil medium such as the IF AT cannot. Similarly, computer-assisted training enables the repeated presentation of questions to learners which, when coupled to branching, may permit a more thorough assessment of the effects of feedback on learning than that provided by the IF AT. It would be of interest in future studies to directly compare the acquisition and retention of course materials, in and outside of the classroom, when feedback is presented using either the IF AT or a computer.

A consistent finding across our studies is the failure to support the delay-retention effect. We have consistently observed that immediate, rather than delayed feedback, results in the greatest increases in retention, confidence, and ability to identify initially-incorrect and correct responses, and the greatest decreases in perseverative incorrect responding. In the present studies these outcomes were

sensitive to the wording of test items, with lower rates of retention, error correction, initial response identification (correct and incorrect), and confidence in identification accuracy observed for items presented on classroom and final examinations with modified wording. It is possible that immediate feedback increases the depth at which corrective information is processed, as demonstrated by Lhyle and Kulhavy (1987).

It is important to contextualize the present discussion of immediate and delayed feedback within the tasks presented to the participants: the selection of correct answers on multiple-choice examinations. Responding to multiple-choice items requires a combination of recognition and discrimination, with recognition more central when test items repeated across examinations are presented with the same wording, and discrimination more central when test items are presented in a modified wording. The task of the learner during multiple-choice testing is to resolve one test item at a time, and while related topics may be presented in adjacent questions, each test item is a discrete unit. Performance on discrete tasks should be facilitated by the provision of immediate, self-corrective feedback, although there may be some learning tasks, such as concept learning, that might be facilitated when feedback is delayed rather than immediate. We are currently examining the effects of immediate and delayed feedback on the acquisition and demonstration of social skills by developmentally-delayed adolescents.

Proponents of immediate feedback theorize that the earlier corrective information is provided, the more likely it is that efficient retention will result (Phye & Andre, 1989). The superiority of immediate feedback has been robustly demonstrated for the acquisition of verbal materials (Ammons, 1956) and motor skills (Anderson, Magill, & Seklya, 2001; Brosvic & Cohen, 1988). The results

observed when immediate feedback is delivered through the IF AT are similar to those observed in numerous studies despite considerable variation in the definition of immediate feedback (i.e., feedback provided immediately after a response to feedback provided by the end of the day), the use or nonuse of an answer-until-correct process, the provision of partial (e.g., 50% of test items) to complete feedback, and differing stimulus materials (fictional materials to classroom concepts). In spite of these considerable differences, the provision of immediate feedback improved the formation of classes and class relations by undergraduates (Adams & Fields, 1999), increased undergraduates' knowledge of biological concepts presented using the PLATO system (Dempsey & Litchfield, 1993), enhanced the test-taking performance of fifth graders (Hanna, 1976) and seventh and eight-grade students completing an achievement test (Hanna & Long, 1979), and increased the general course performance of junior high school and college students (Beeson, 1973).

The present findings are of particular interest as the comparisons in Study 2 were made within-subjects and after experience with each response format, and in both studies, the control procedure was one of the most commonly used student assessment formats: the Scantron form. These outcomes have considerable implications for how practice tests, such as publisher-supplied study guides and web-based testing, could be structured in order to maximize learning and retention. Collectively, the results of Studies 1 and 2 demonstrate that immediate feedback increases retention, accuracy at identifying initial responses, and confidence in response identifications while simultaneously reducing perseverative incorrect responding.

References

Adams, B. J., & Fields, L. (1999) Effects of unreinforced conditional selection training, multiple negative comparison training, and feedback on equivalence class formation. The Psychological Record, 49, 685-702.

Anderson, D. I., Magill, R. A., & Seklya, H. (2001) Motor learning as a function of KR schedule and characteristics of task-intrinsic feedback. Journal of Motor Behavior, 33, 59-66.

Ammons, R. B. (1956) Effects of knowledge of performance: A survey of theoretical formulation. Journal of General Psychology, 54, 279-299.

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

Brackbill, Y., Bravos, A., & Starr, R. H. (1962). Delay-improved retention of a difficult task. Journal of Comparative and Physiological Psychology, 55, 947-952.

Brosvic, G. M., & Cohen, B. (1988) The horizontal-vertical illusion and knowledge of results. Perceptual and Motor Skills, 67, 463-469.

Dempsey, J. V., & Litchfield, B. C. (1993) Feedback, retention, discrimination error, and feedback study time. Journal of Research on Computing in Education, 25, 303-326.

Dihoff, R. E., Brosvic, G. M., & Epstein, M. L. (In press) The role of feedback during academic testing: The delay retention effect revisited. The Psychological Record.

Epstein, M. L., & Brosvic, G. M. (2002) Students prefer the immediate feedback assessment technique. Perceptual and Motor Skills, 90, 1136-1138.

Epstein, M. L., Brosvic, G. M., Dihoff, R. E., Lazarus, A. D., & Costner, K. L. (In press) Effectiveness of feedback during the testing of preschool children, elementary school children, and adolescents with developmental delays. The Psychological Record.

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

Epstein, M. L., Lazarus, A. D., Calvano, T. B., Matthews, K. A., Hendel, R. A., Epstein, B. B., & Brosvic, G. M. (2002) Immediate feedback assessment technique promotes learning and corrects inaccurate first responses. The Psychological Record, 52, 187-201.

Hanna, G. H. (1976) Effects of total and partial feedback in multiple-choice testing upon learning. The Journal of Educational Research, 69, 202-205.

Hanna, G. M., & Long, C. A. (1979) Effect of answer until correct testing on reliability. Perceptual and Motor Skills, 49, 464-466.

Kritch, K. M., & Bostow, D. E. (1998) Degree of constructed-response interaction in computer-based programmed instruction. Journal of Applied Behavior Analysis, 31, 387-398.

Kulhavy, R. W., & Anderson, R. C. (1972) Delay-retention effects with multiple choice tests. Journal of Educational Psychology, 63, 505-512.

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

Miller, M. L., & Malott, R. W. (1997) The importance of overt responding in programmed instruction even with added incentives for learning. Journal of Behavioral Education, 7, 497-503.

O'Neill, M., Rasor, R. A., & Bartz, W. R. (1976) Immediate retention of objective test answers as a function of feedback complexity. The Journal of Educational Research, 70, 72-75.

Pressey, S. L. (1926) A simple device which gives tests and scores and teaches. School and Society, 23, 373-376.

Robin, A. L. (1978) The timing of feedback in personalized instruction. Journal of Personalized Instruction, 3, 81-87.

Skinner, B. F. (1958) Teaching machines. Science, 128, 969-977.

Surber, J. R., & Anderson, A. C. (1975) Delay-retention effect in natural classroom settings. Journal of Educational Psychology, 67, 170-173.

Tudor, R. M. (1995) Isolating the effects of active responding in computer-based instruction. Journal of Applied Behavior Analysis, 28, 343-344.

Table 1

Conditional Probability (in percentages) of Responding on Items With Identical Wording Across Administrations in Study 1

	IF AT	FB - 3/3	FB - 0/3	Scantron
Practice Test to Classroom Examination				
Correct / Correct	76.34	53.56	32.34	39.48
Incorrect / Correct	51.05	30.76	12.04	16.33
Correct / Incorrect	23.66	46.44	67.66	60.53
Incorrect / Incorrect	48.95	69.24	87.96	83.67
Practice Test to Final Examination				
Correct / Correct	71.87	48.06	27.78	34.41
Incorrect / Correct	47.91	29.68	10.54	14.53
Correct / Incorrect	28.13	51.94	72.22	65.59
Incorrect / Incorrect	52.09	70.32	89.46	85.47
Classroom Examination to Final Examination				
Correct / Correct	73.04	57.03	25.17	31.82
Incorrect / Correct	57.72	34.68	18.89	20.29
Correct / Incorrect	26.94	42.97	74.83	67.18
Incorrect / Incorrect	42.28	65.32	81.11	79.71

Table 2

Conditional Probability (in percentages) of Responding on Items With Modified Wording Across Administrations in Study 1

	Practice Test to Classroom Examination			
	IF AT	FB - 3/3	FB - 0/3	Scantron
Correct / Correct	48.28	36.02	24.34	26.76
Incorrect / Correct	39.03	25.66	11.08	14.32
Correct / Incorrect	51.72	63.98	75.66	73.24
Incorrect / Incorrect	60.97	74.34	88.92	85.68
	Practice Test to Final Examination			
	IF AT	FB - 3/3	FB - 0/3	Scantron
Correct / Correct	43.69	33.12	19.01	25.38
Incorrect / Correct	37.72	28.45	17.78	20.19
Correct / Incorrect	56.31	66.88	80.99	74.62
Incorrect / Incorrect	62.28	71.55	82.22	79.81
	Classroom Examination to Final Examination			
	IF AT	FB - 3/3	FB - 0/3	Scantron
Correct / Correct	45.99	30.27	13.38	10.98
Incorrect / Correct	39.68	25.88	16.04	11.34
Correct / Incorrect	54.01	69.73	86.62	89.02
Incorrect / Incorrect	60.32	74.12	83.96	88.66

Table 3

Post-Test Measures Assessing Perceptions as a Function of The Number of Practice Tests During Which Feedback Was Provided in Study 1

		IF AT	FB - 3/3	FB - 0/3	Scantron
Test	M	3.09	3.11	3.15	2.93
Anxiety	SD	1.04	1.12	1.23	1.19
Test		3.09	2.98	3.15	2.93
Anxiety		1.04	1.05	1.23	1.19
Time		3.32	3.19	3.15	3.17
Requirements		1.28	1.44	1.33	1.01
Satisfaction With		4.14	4.19	2.94	2.78
Response Format		1.56	1.78	1.02	0.78
Clarity of Response		4.23	4.45	2.95	3.01
Requirements		1.41	1.24	1.03	1.20
Benefits of		4.55	4.28	2.89	2.64
Testing		0.78	1.56	1.23	0.66
Desirability of		4.42	4.33	2.85	2.73
Response Format		1.04	1.22	1.23	0.89

Table 4

Conditional Probability (in percentages) of Responding on Items With Identical Wording Across Administrations in Study 2

	Practice Test to Classroom Examination			
	IF AT	End of Test Feedback	Delayed Feedback	Scantron
Correct / Correct	67.03	38.77	35.87	20.24
Incorrect / Correct	45.88	25.78	28.81	22.28
Correct / Incorrect	32.97	62.23	64.13	79.76
Incorrect / Incorrect	54.12	74.22	71.19	77.72
	Practice Test to Final Examination			
	IF AT	End of Test Feedback	Delayed Feedback	Scantron
Correct / Correct	70.88	35.08	31.74	18.07
Incorrect / Correct	46.54	22.99	23.34	11.62
Correct / Incorrect	29.12	64.92	68.26	81.93
Incorrect / Incorrect	53.46	77.01	76.56	88.38
	Classroom Examination to Final Examination			
	IF AT	End of Test Feedback	Delayed Feedback	Scantron
Correct / Correct	63.27	28.18	25.73	15.48
Incorrect / Correct	50.45	17.27	21.19	09.85
Correct / Incorrect	36.73	71.82	74.27	84.52
Incorrect / Incorrect	49.55	82.73	78.81	90.15

Table 5

Conditional Probability (in percentages) of Responding on Items With Modified Wording Across Administrations in Study 2

	Practice Test to Classroom Examination			
	IF AT	End of Test Feedback	Delayed Feedback	Scantron
Correct / Correct	46.02	28.49	24.07	12.99
Incorrect / Correct	37.49	20.43	18.79	08.81
Correct / Incorrect	53.98	71.51	65.93	87.01
Incorrect / Incorrect	62.51	79.57	81.21	91.19
	Practice Test to Final Examination			
	IF AT	End of Test Feedback	Delayed Feedback	Scantron
Correct / Correct	42.62	25.41	21.98	10.72
Incorrect / Correct	31.18	18.87	24.11	11.17
Correct / Incorrect	57.38	74.59	78.02	89.28
Incorrect / Incorrect	68.82	81.13	75.89	88.73
	Classroom Examination to Final Examination			
	IF AT	End of Test Feedback	Delayed Feedback	Scantron
Correct / Correct	47.85	26.38	27.05	14.88
Incorrect / Correct	41.58	21.72	26.74	12.09
Correct / Incorrect	52.15	73.62	72.95	85.12
Incorrect / Incorrect	58.42	78.28	73.26	87.91

Table 6

Post-Test Measures Assessing Perceptions as a Function of The Timing of Feedback in Study 2

		IF AT	End of Test Feedback	Delayed Feedback	Scantron
Test Anxiety	M	2.87	2.58	2.75	2.81
	SD	1.25	1.04	1.15	1.35
Time Requirements		2.68	2.81	2.99	3.02
		1.11	0.58	0.78	1.58
Satisfaction With Response Format		4.38	3.02	3.21	2.75
		1.07	1.77	1.18	1.44
Clarity of Response Requirements		4.11	3.15	2.87	2.98
		1.33	1.08	1.45	1.37
Benefits of Testing		4.65	3.47	3.35	2.87
		1.25	1.44	1.08	1.65
Desirability of Response Format		4.28	3.02	2.87	2.69
		1.23	1.75	1.12	1.73

Figures

1. Sample portion of the Immediate Feedback Assessment Technique (IF AT) form. Patent is held by E3 Corporation.
2. Scores on classroom examinations and the final examination as a function of the number of practice tests with feedback. [Legend: IF AT - closed circle; Scantron form - open circle.]
3. Percentage of correctly-answered items with identical wording across administration as a function of the number of practice tests with feedback. [Legend: IF AT - closed circle; Scantron form - open circle.]
4. Percentage of correctly-answered items with modified wording across administration as a function of the number of practice tests with feedback. [Legend: IF AT - closed circle; Scantron form - open circle.]
5. Mean identification accuracy for correctly identified responses that were initially-correct on the final examination as a function of number of practice tests with feedback.
6. Mean identification accuracy for correctly identified responses that were initially-incorrect on the final examination as a function of number of practice tests with feedback.
7. Mean confidence ratings when identifying initial responses on the final examination as a function of the number of practice tests with feedback and item wording.
8. Scores on classroom examinations and the final examination as a function of the timing of feedback. [Legend: IF AT - closed circle; End of Test Feedback - open square; Delayed Feedback - open circle; Scantron form - open diamond.]

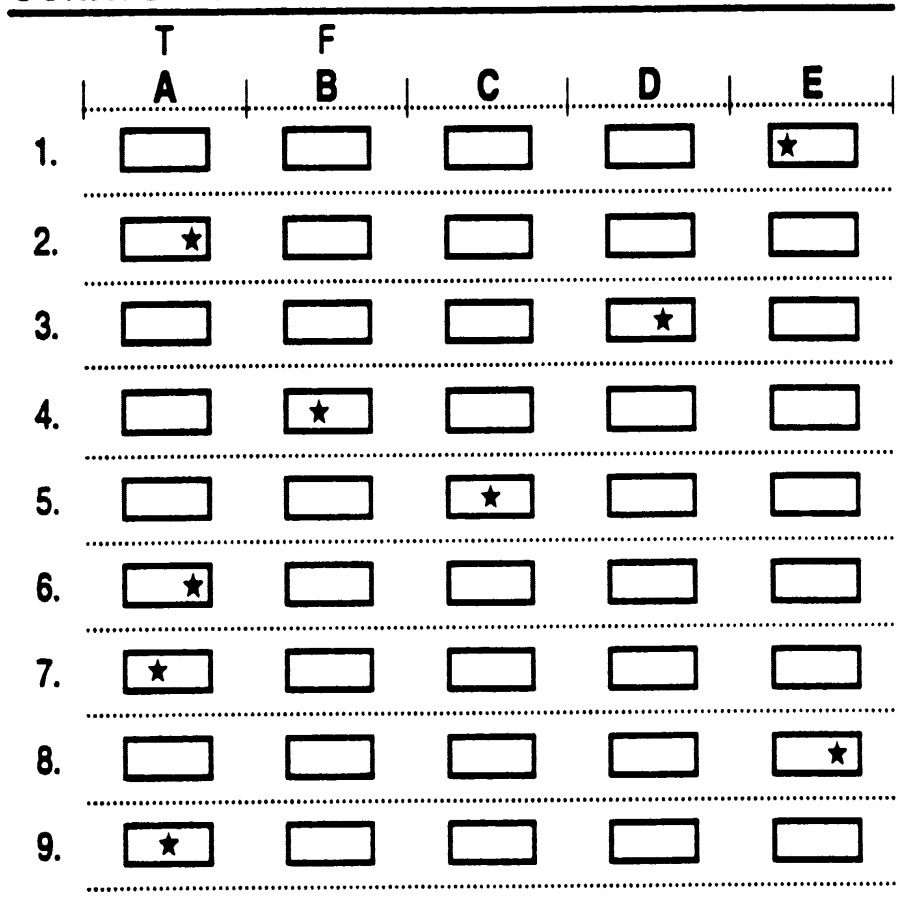
9. Percentage of correctly-answered items with identical wording across administration as a function of the timing of feedback. [Legend: IF AT - closed circle; End of Test Feedback - open square; Delayed Feedback - closed square; Scantron form - open circle.]
10. Percentage of correctly-answered items with modified wording across administration as a function of the timing of feedback. [Legend: IF AT - closed circle; End of Test Feedback - open square; Delayed Feedback - closed square; Scantron form - open circle.]
11. Mean identification accuracy for correctly identified responses that were initially-correct on the final examination as a function of the timing of feedback.
12. Mean identification accuracy for correctly identified responses that were initially-incorrect on the final examination as a function of the timing of feedback.
13. Mean confidence ratings when identifying initial responses on the final examination as a function of the timing of feedback.

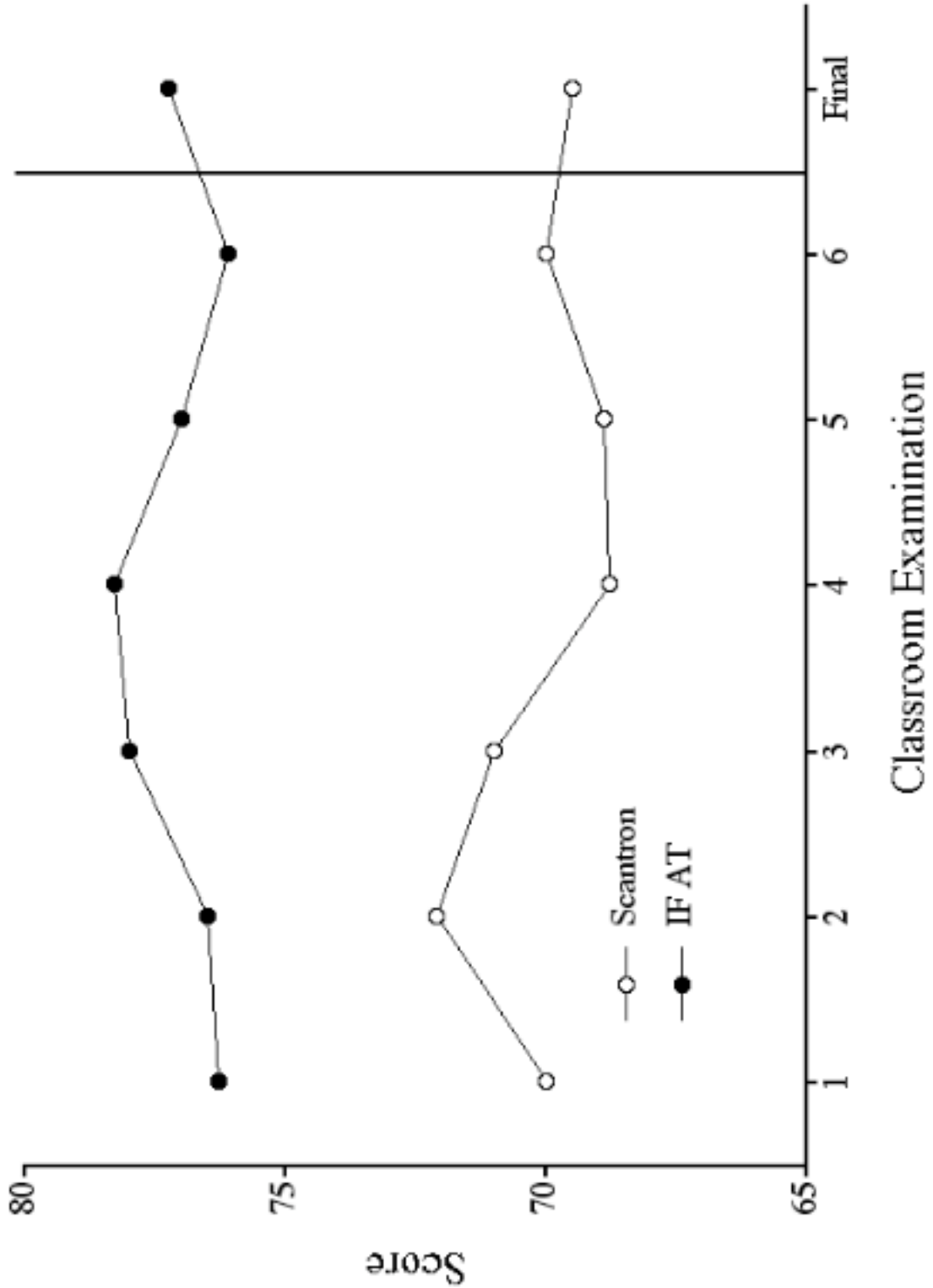
IMMEDIATE FEEDBACK ASSESSMENT TECHNIQUE (IF AT)

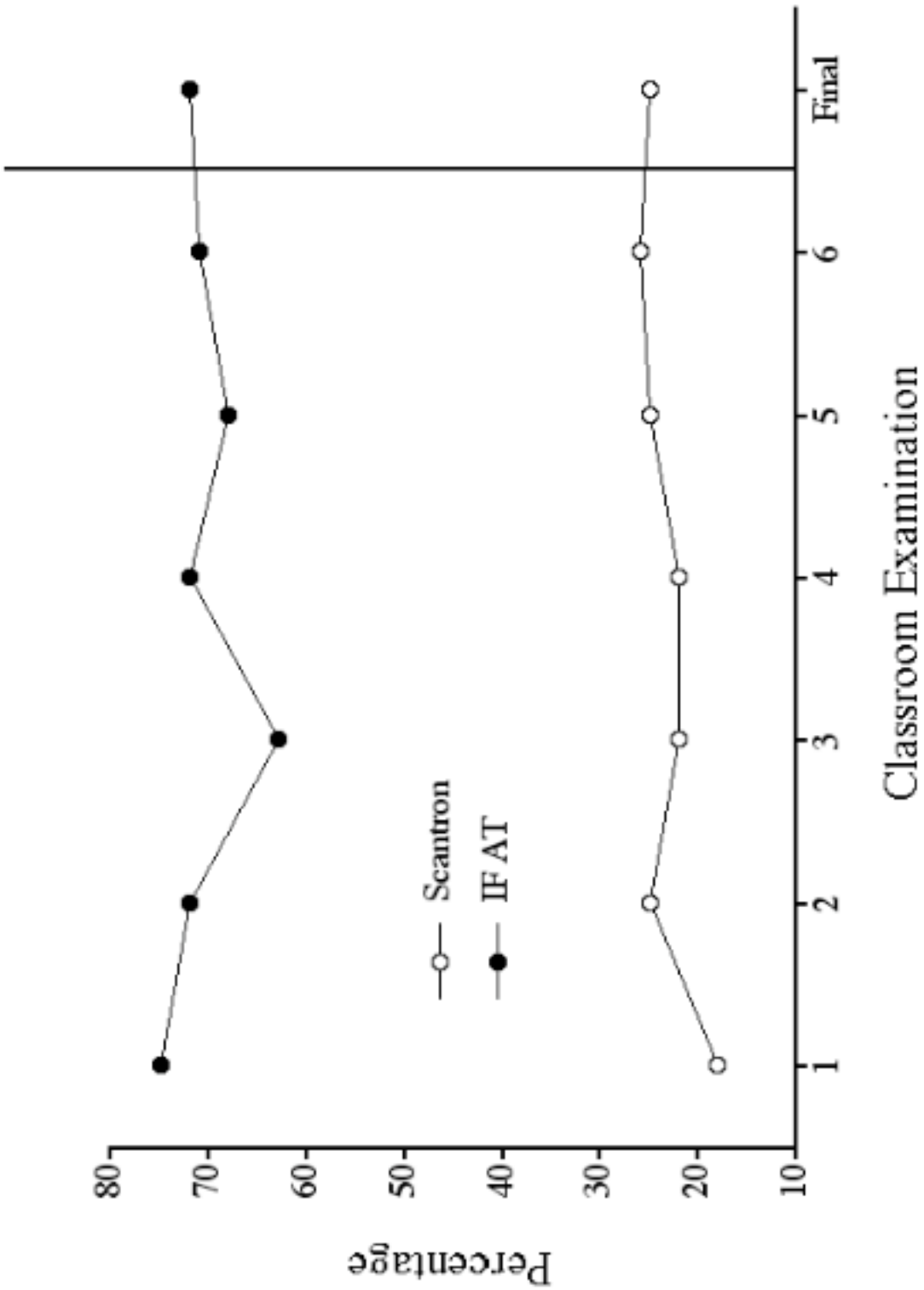
Name _____ Test # _____

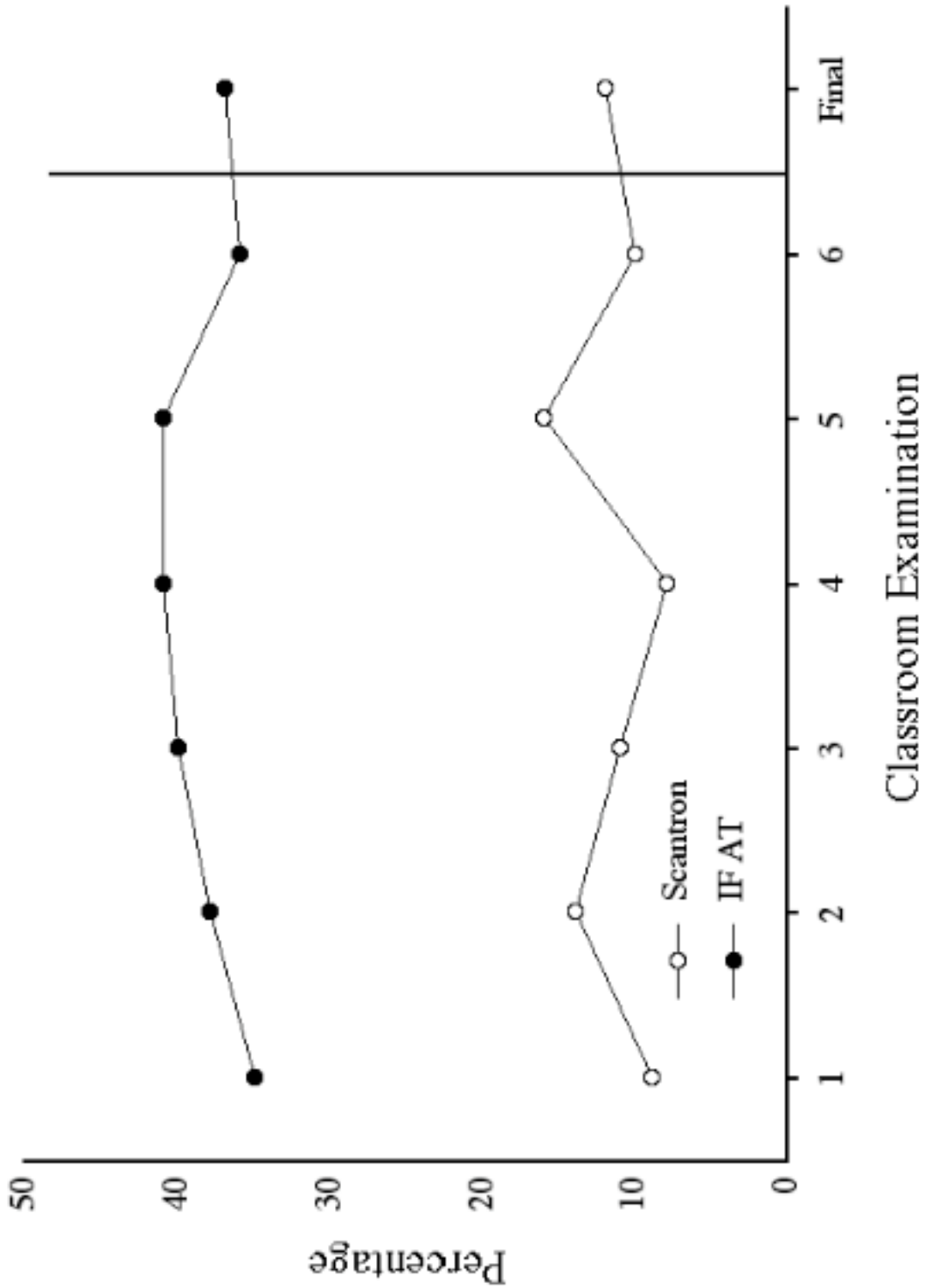
Subject _____ Score _____

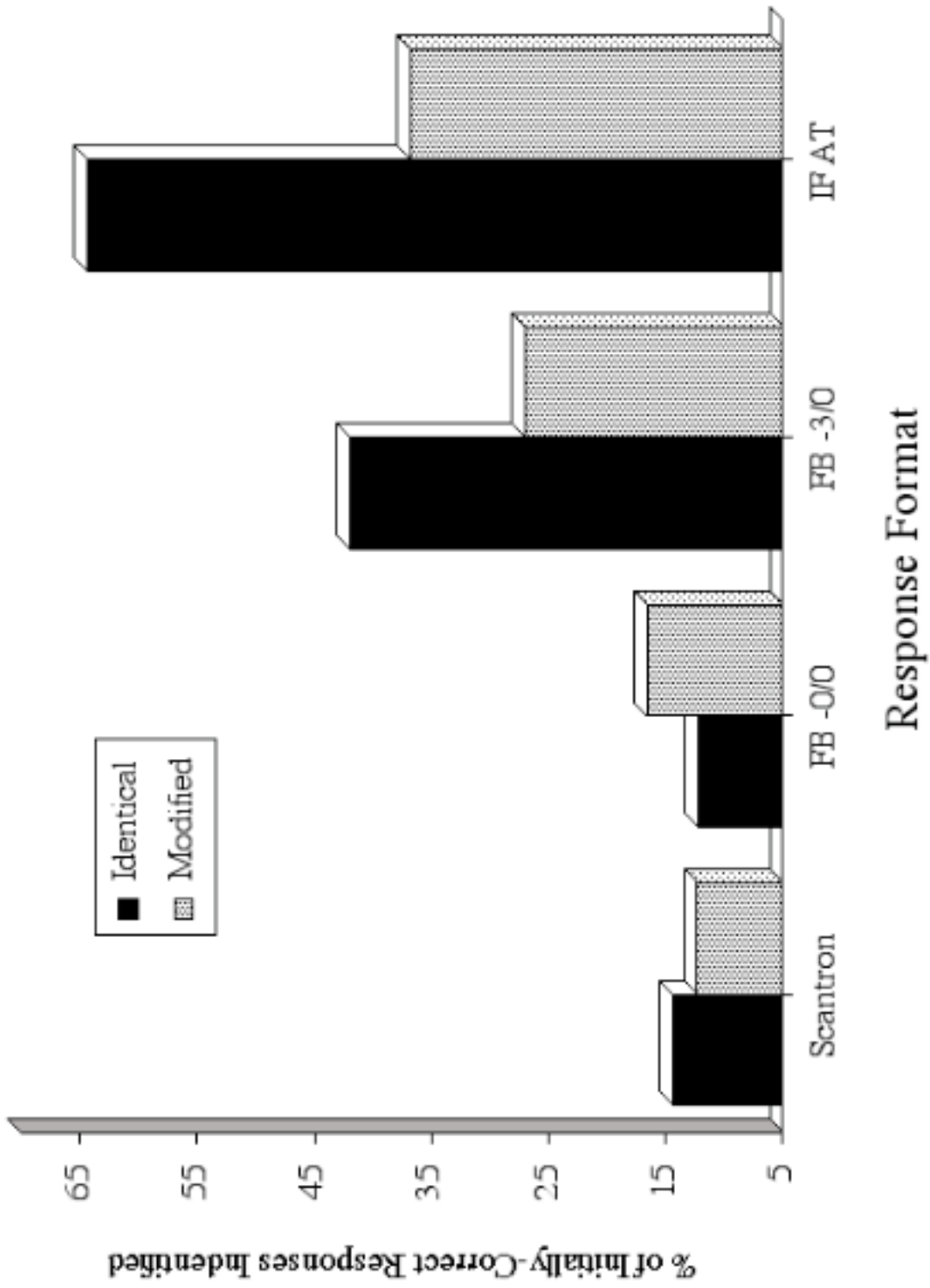
SCRATCH OFF COVERING TO EXPOSE ANSWER

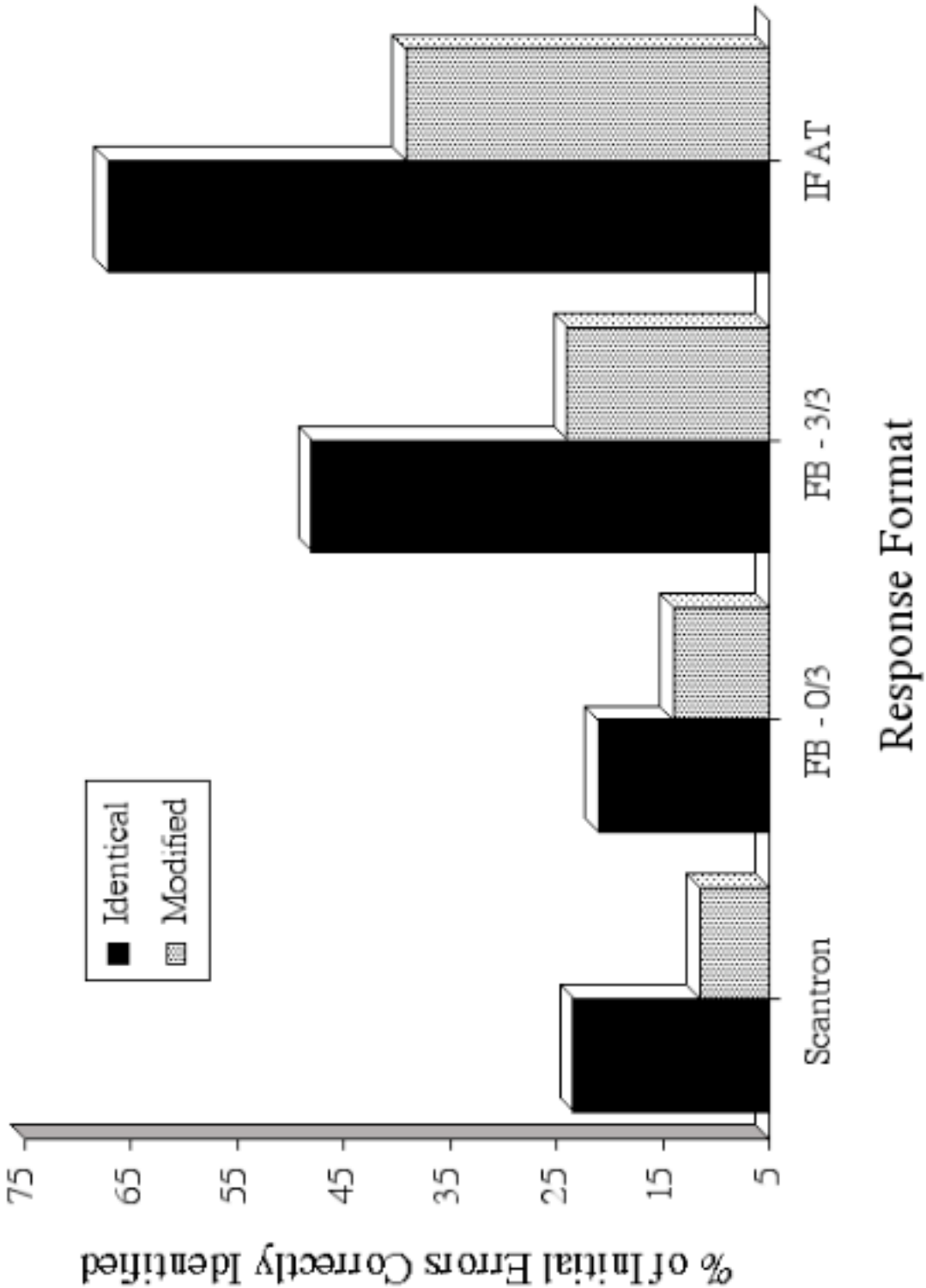


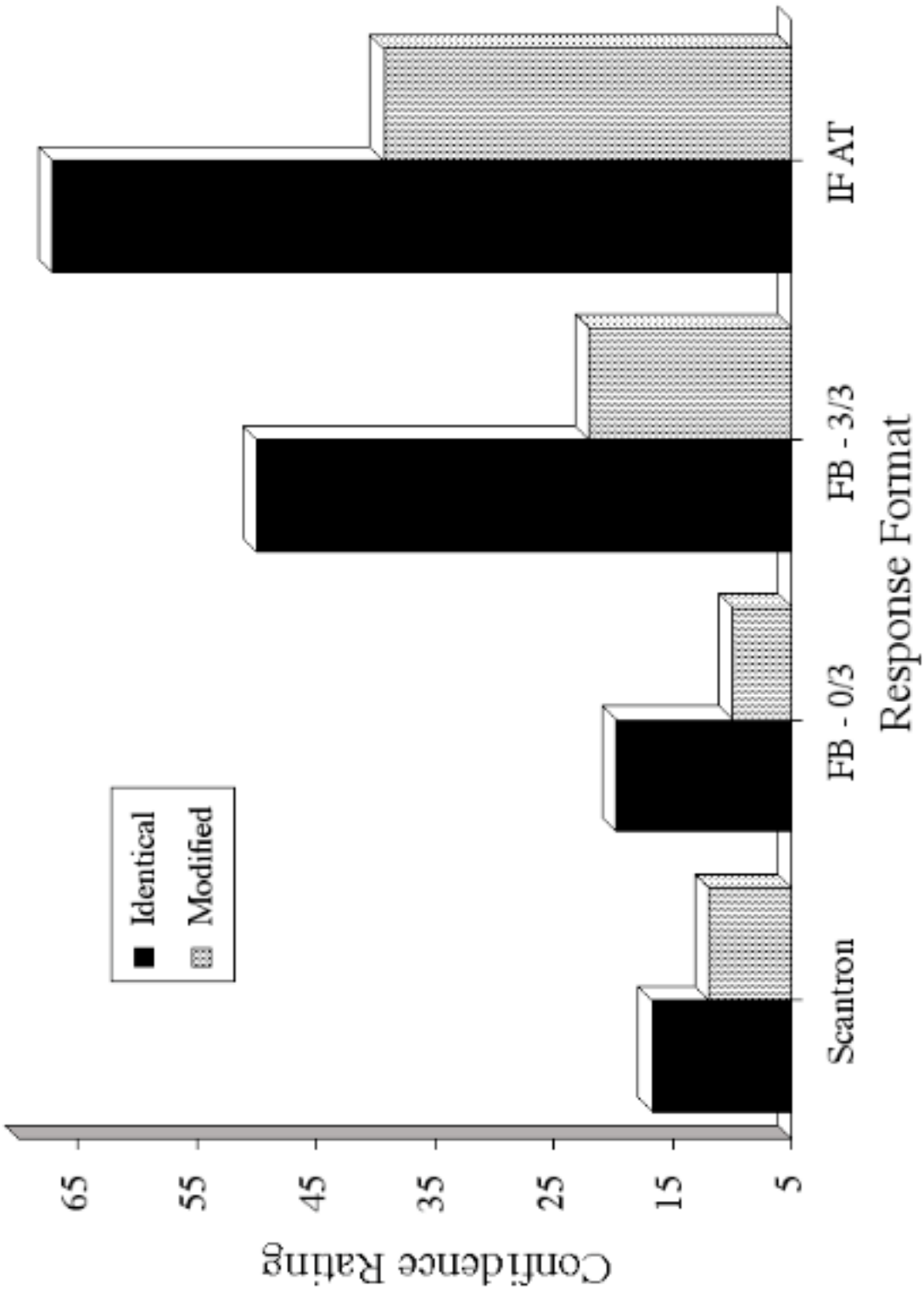

Score
80
75
70
65
1
2
3
4
5
6
Final
Classroom Examination
Scantron
IF AT

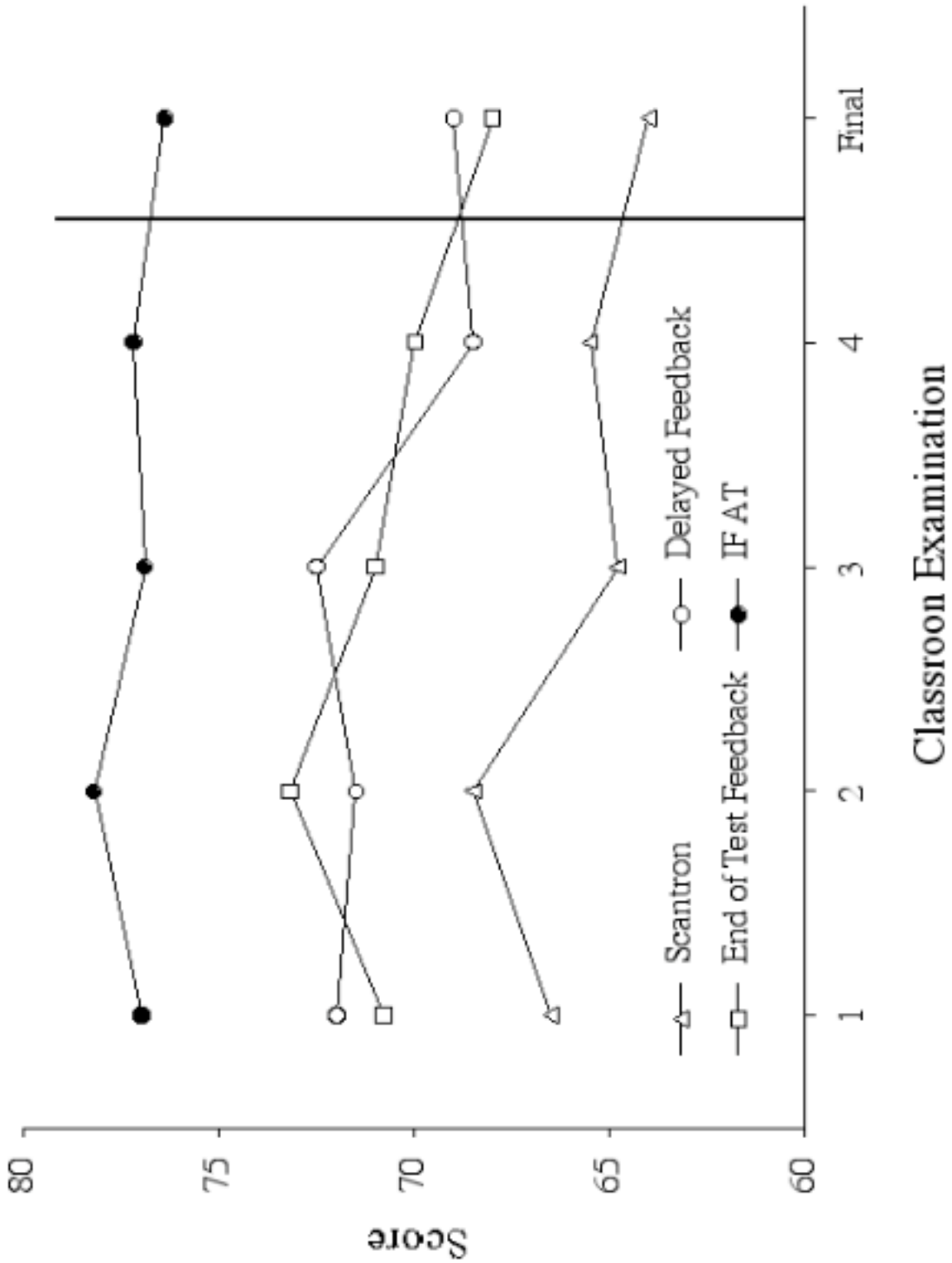

Percentage
80
70
60
50
40
30
20
10
Scantron
IF AT
1
2
3
4
5
6
Final
Classroom Examination

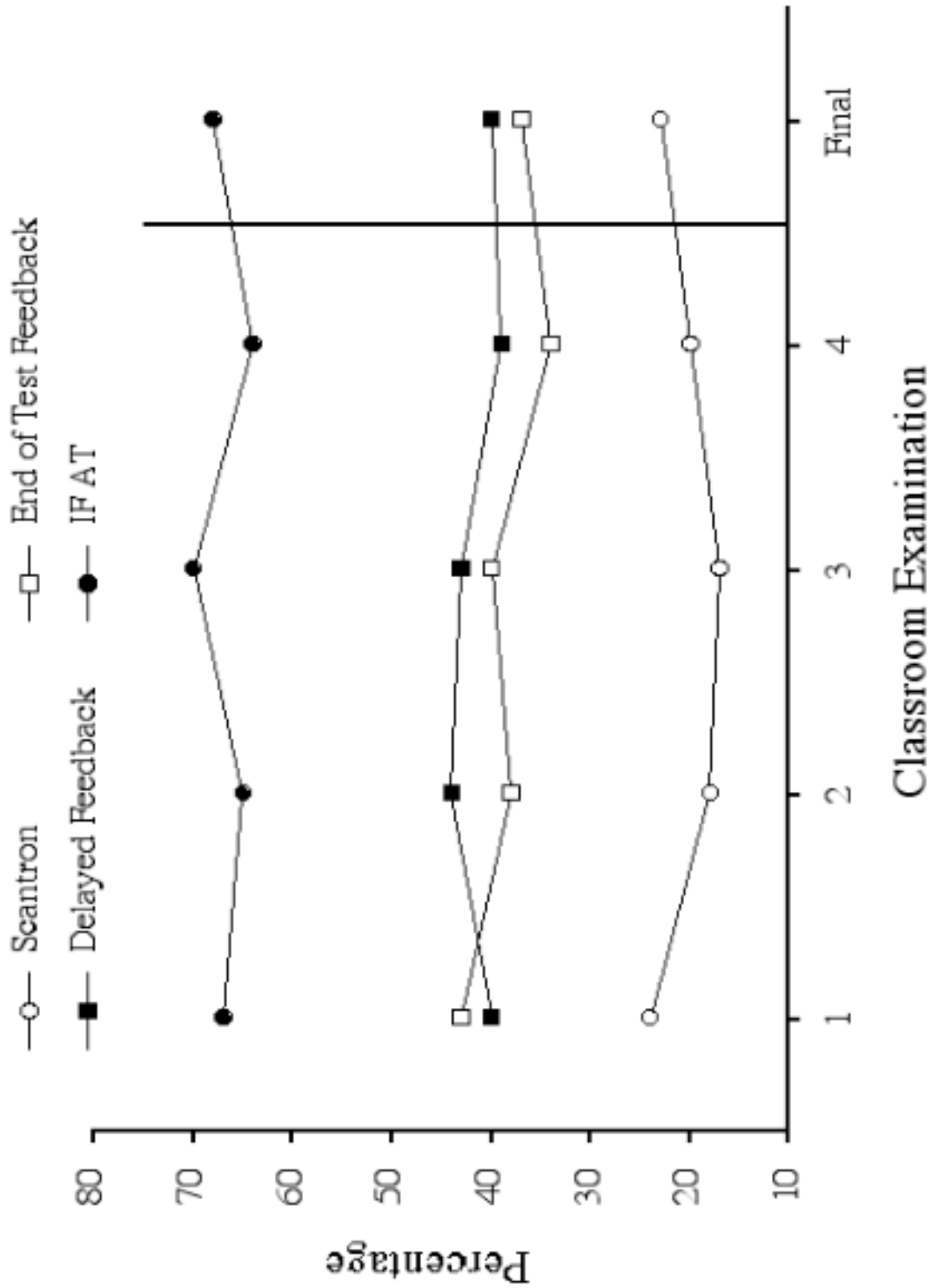

Percentage
50
40
30
20
10
0
1
2
3
4
5
6
Final
Classroom Examination
Scantron
IF AT

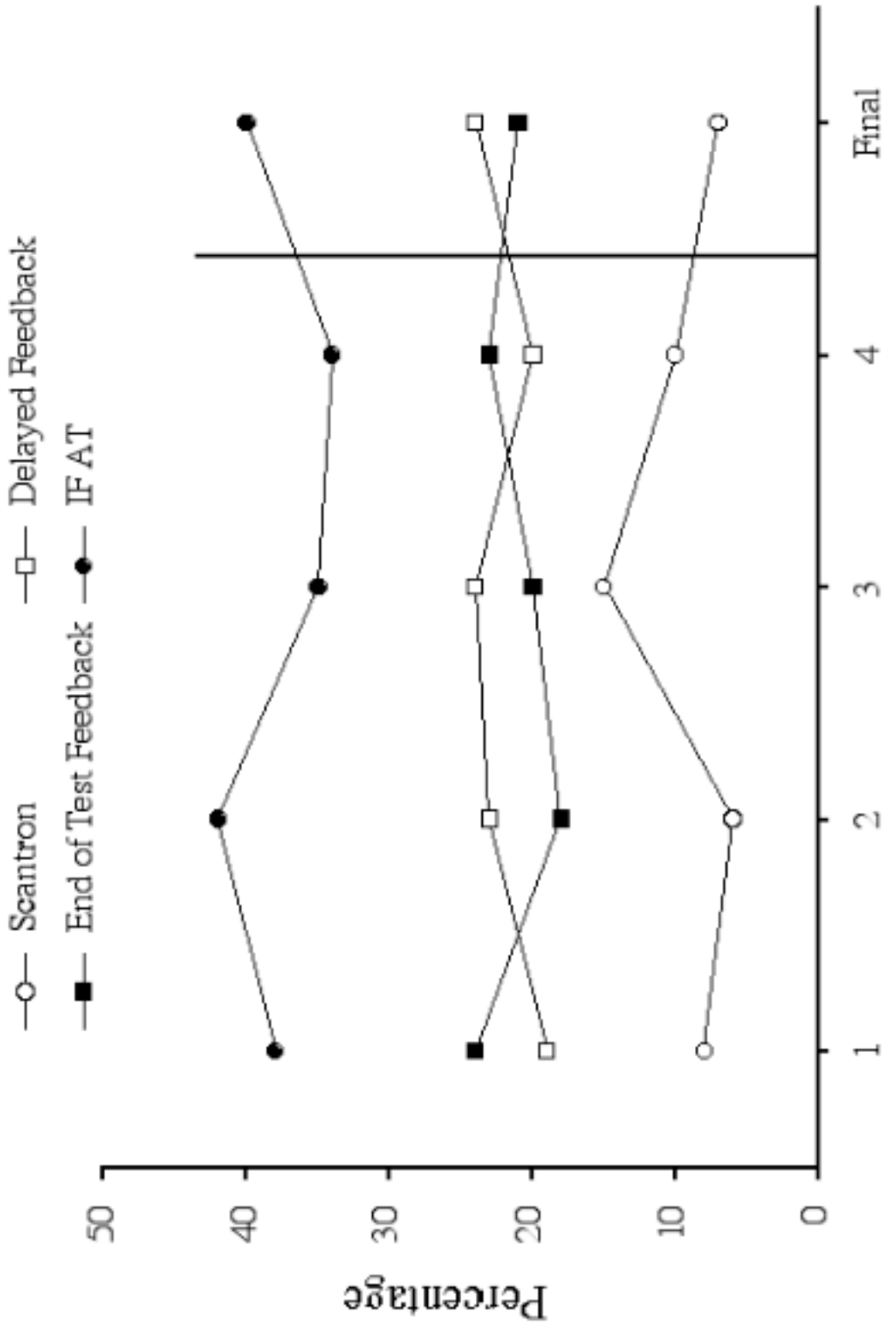

% of Initially-Correct Responses Identified
65
55
45
35
25
15
5
Identical
Modified
Scantron
FB -0/0
FB -3/0
IF AT
Response Format

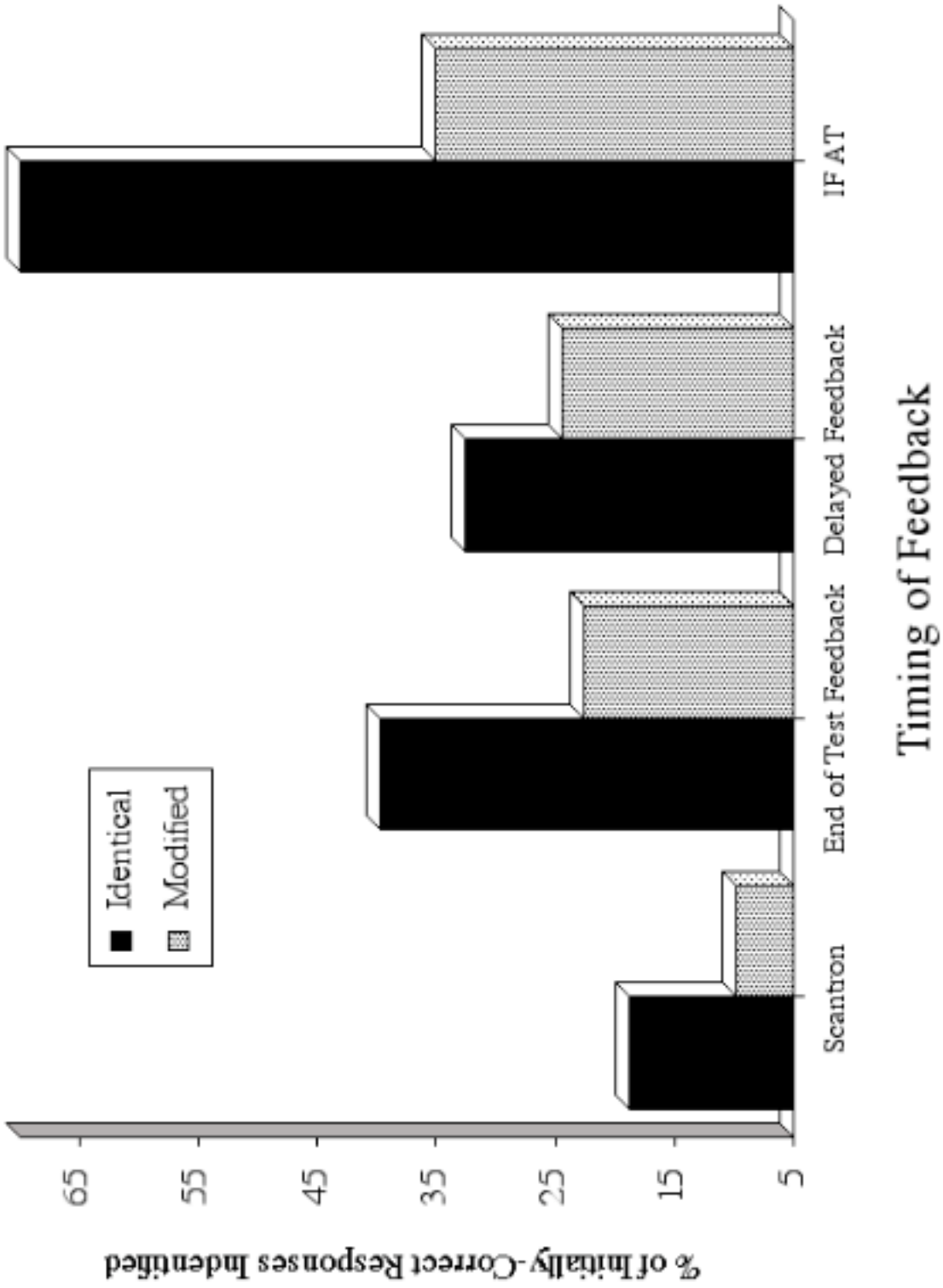

% of Initial Errors Correctly Identified
75
65
55
45
35
25
15
5
Identical
Modified
Scantron
FB - 0/3
FB - 3/3
IF AT
Response Format

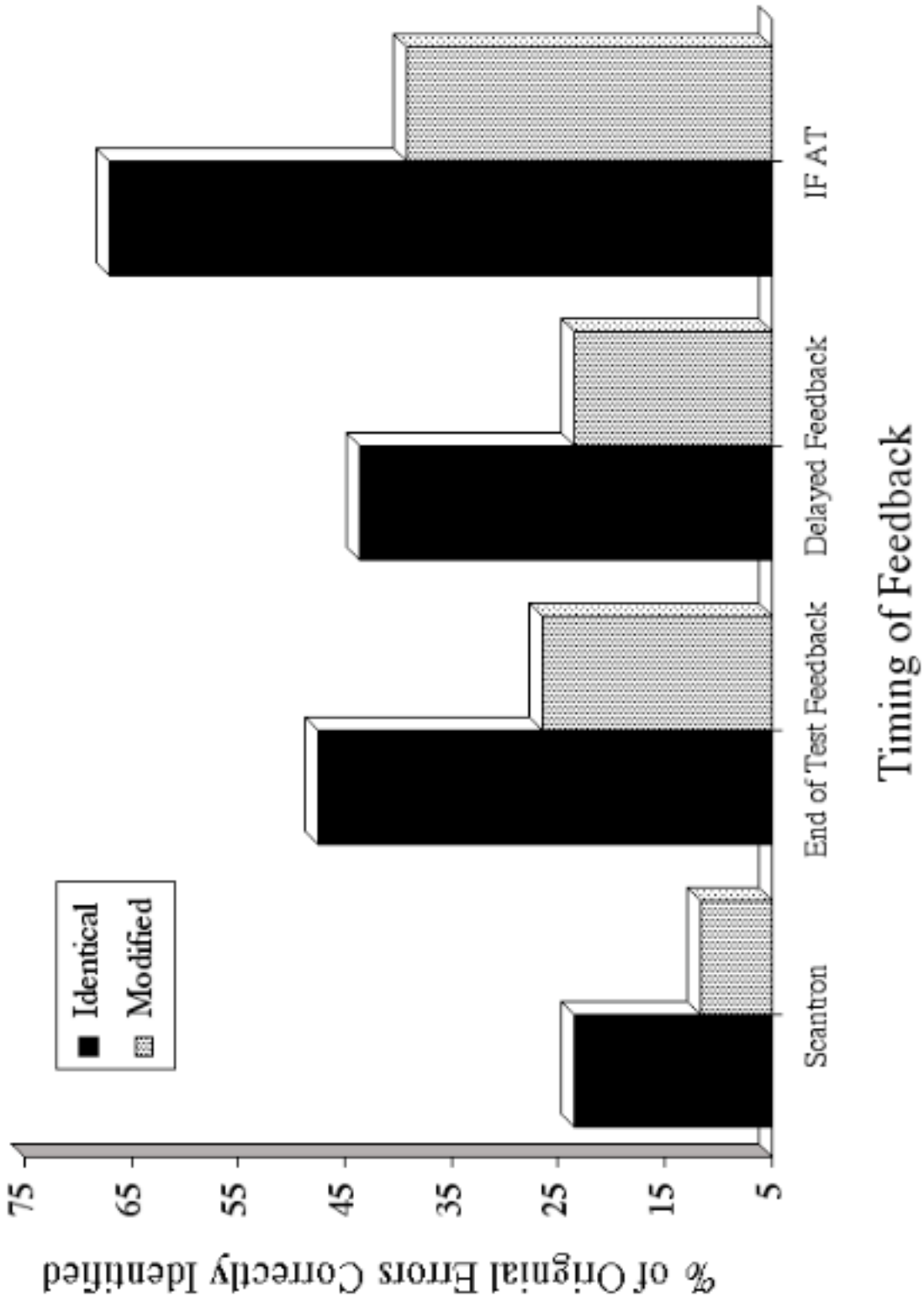

Confidence Rating
65
55
45
35
25
15
5
Identical
Modified
Scantron
FB - 0/3
FB - 3/3
IF AT
Response Format

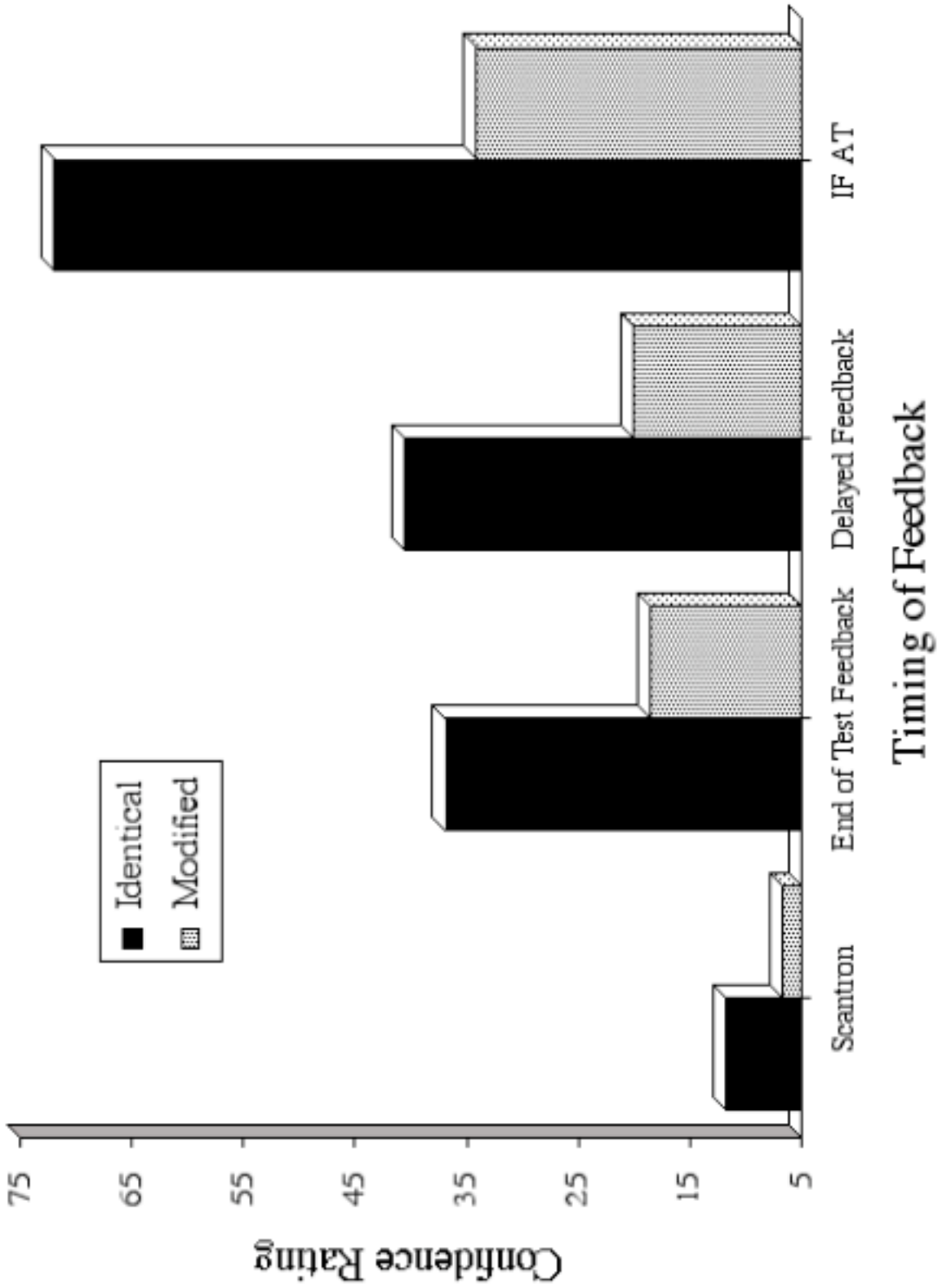

Score
80
75
70
65
60
1
2
3
4
Final
Classroon Examination
Scantron
Delayed Feedback
End of Test Feedback
IF AT


Scantron
End of Test Feedback
Delayed Feedback
IF AT
Percentage
80
70
60
50
40
30
20
10
1
2
3
4
Final
Classroom Examination


Scantron
Delayed Feedback
End of Test Feedback
IF AT
Percentage
50
40
30
20
10
0
1
2
3
4
Final
Classroom Examination


% of Initially-Correct Responses Identified
65
55
45
35
25
15
5
Identical
Modified
Scantron
End of Test Feedback
Delayed Feedback
IF AT
Timing of Feedback


% of Original Errors Correctly Identified
75
65
55
45
35
25
15
5
Identical
Modified
Scantron
End of Test Feedback
Delayed Feedback
IF AT
Timing of Feedback


Identical
Modified
75
65
55
45
35
25
15
5
Confidence Rating
Scanton
End of Test Feedback
Delayed Feedback
IF AT
Timing of Feedback