JOURNAL OF APPLIED BEHAVIOR ANALYSIS 1974, 7, 61-69 NUMBER 1 (SPRING 1974)

THE EFFECTS OF MASTERY CRITERIA AND ASSIGNMENT LENGTH ON COLLEGE-STUDENT TEST PERFORMANCE[1]

GEORGE SEMB

UNIVERSITY OF KANSAS

This study analyzed the function of two components of a personalized instruction course —mastery criteria for passing a test and assignment length. A high mastery criterion (100% correct) and short assignments produced better test performance than either a low mastery criterion (60% correct) or long assignments (four short assignments combined) on both study question items that students had in their possession and probe items that were not available to students in advance.

Several investigators (Alba and Pennypacker, 1972; Born, Gledhill, and Davis, 1972; McMichael and Corey, 1969; Sheppard and MacDermot, 1970) have compared examination performance following instruction with traditional lecture procedures *versus* personalized instruction procedures. The general results indicate that personalized instruction produces superior examination performance and higher student ratings than more traditional lecture-discussion procedures.

Recently, investigators have begun to separate and analyze the components of personalized instruction originally outlined by Keller (1968). For example, Farmer, Lachter, Blaustein, and Cole (1972) studied the role of proctoring in personalized instruction. They found that proctored students performed better on a final examination than non-proctored students, and that increased proctoring accelerated rate of progress through the course. Semb, Hopkins, and Hursh (1973) investigated the effects of study questions on examination performance. They found that students performed better on study question items than on non-study question items on both short quizzes and comprehensive hour examinations. Thus, two components of the personalized instruction package—student proctors and study questions—have been isolated and investigated, and both appear to be functionally related to student test performance. The purpose of the present study was to analyze the effects of two additional components of personalized instruction: (1) mastery criteria and (2) assignment length.

In one of his early articles on personalized instruction, Keller (1968) defined mastery as perfect performance (100% correct) on short unit quizzes. Other investigators have used different definitions of mastery. For example, Semb *et al.* (1973) defined mastery as 90% correct on short unit quizzes. In a related experiment, Johnston and O'Neill (1973) evaluated student rates of correct and incorrect responding under several different mastery conditions (High, Medium, Low, and None). Students with no criterion on performance exhibited the lowest exam scores. Furthermore, performance levels were directly related to the parametric value of the mastery criterion. The present study systematically extended the Johnston and O'Neill experiment. Although the question is similar, course content, experimental design, and response measures are different.

Keller (1968) also advocated the use of short unit assignments. The empirical question is: do short unit assignments produce different effects on performance than long assignments, when a long assignment is defined as the com-

[1]Portions of this research were supported by the University of Kansas Computation Center. Reprints may be obtained from George Semb, Department of Human Development, University of Kansas, Lawrence, Kansas 66044.

bination of several short assignments? If, for example, long assignments produce examination performance that is equivalent to or better than short assignments, instructors could potentially save a great deal of time and money (*e.g.*, preparation time, proctor time, and printing costs). If, on the other hand, long assignments lead to inferior performance, instructors might be well advised to take large segments of material and divide them into short units, as originally proposed by Keller (1968).

METHOD

Subjects and Course Personnel

Students enrolled in an introductory child development course at the University of Kansas served as subjects. One instructor, two teaching assistants, and 10 undergraduate proctors managed the course. Students were randomly assigned to one of two groups during enrollment. Of the 51 students who enrolled in Group 1, five withdrew from the course, leaving 46 students, with whom the present study was conducted. Of the 52 students who enrolled in Group 2, nine withdrew, leaving a total of 43. Five students who withdrew from the course (two from Group 1 and three from Group 2) indicated that they were unable to maintain the minimum progress criterion set by the instructor. The remaining nine students withdrew from the course during the first two weeks of the semester when no explanation for a withdrawal was required.

Procedure

The course was divided into four major content parts; each part was further subdivided into four units. A unit consisted of a 30-page reading assignment[2] and 16 short-answer essay study questions. The course was both student-paced and instructor-paced, in that the student could work as fast as he wished, but had to maintain a minimum rate of progress (on the average of one quiz passed every 1.5 days) or withdraw from the course (Miller, Weaver, and Semb, 1974). The semester lasted for six weeks, or 30 class days.

To earn a grade of A, students had to pass a series of quizzes. They could take each quiz (different forms) as many times as necessary to meet the predetermined mastery criterion. Since students had to maintain a minimum rate of progress (defined as mastery of successive quizzes) throughout the course, only the grades of A and W (Withdrawal) were available.

Both groups met in different sections of the same room, and both were informed of the procedural differences between groups. Quizzes could be taken at any time between 8:00 a.m. and 12:30 p.m. Monday through Friday. Thus, students could take or re-take several quizzes on the same day.

Content and Review Quizzes

The first four quizzes in each part were defined as "content" quizzes. The fifth was defined as a "review" quiz that covered the preceding four content units. Each content quiz consisted of eight study question items and two probe items. Study items were randomly selected from the student's pool of 16 study questions, and probe items were randomly selected from a pool of nine probe items per unit. Review quizzes consisted of 16 randomly selected study question items (four from each content unit) and four probe items (one from each content unit). Four different forms of each content and review quiz were available.

Study questions were designed to emphasize major points in the readings. Probe items were also designed to cover major points, but more importantly, they generally required students to put together major concepts, ideas, or procedures. For example, a study question asked: "Name the three specific freedoms stressed by Neill and give an example of how one such

[2]Course materials included *Readings in developmental psychology today* (Cramer, 1970), *Elementary principles of behavior* (Whaley and Malott, 1971) and *Human development lecture notes and study guide* (Semb, *unpublished*).

freedom might be fostered at Summerhill." The answer to the question required the student to recall specific, factual information, and he or she was graded on the basis of how closely the answer matched the instructor's answer guide. On the other hand, probe items took the form: "If you train a child to discriminate between a horizontal line and a vertical line, then gradually change the vertical line until it is diagonal, what behavioral procedure are you using? Explain." Here, the student must take a novel example, apply the appropriate label, and explain how the procedure was being implemented. Again, the student's answer was evaluated in terms of how well it matched the instructor's answer guide. It should be pointed out, however, that the *major* difference between the two types of items was whether or not they appeared as part of the student's study questions.

When the student completed a quiz, he or she took it to his proctor, who graded it immediately. The proctor graded each answer as either correct, partially correct, or incorrect. The student could try to justify or explain answers that the proctor marked as questionable or incorrect. If the student's oral explanation satisfied the proctor, the item was marked as correct. If the student's explanation was insufficient, the proctor marked the answer incorrect. Whether or not the student could progress to the next unit depended upon his test performance and the mastery condition in effect.

Experimental Design

The experiment used a within-group reversal design. The manipulations included:

High criterion-short assignment (*HC-SA*). Students took content and review quizzes, all of which had to be passed at a mastery criterion of 100% correct.

Low criterion-short assignment (*LC-SA*). Students took content and review quizzes, all of which had to be mastered at 60% correct or better.

High criterion-long assignment (*HC-LA*). Content quizzes for each unit were eliminated. Students took the review quiz until it was mastered with a score of 100%.

Groups 1 and 2 went through the experimental conditions in different sequences:

Course Part	*1*	*2*	*3*	*4*
Group 1	HC-SA	LC-SA	HC-SA	HC-LA
Group 2	LC-SA	HC-SA	HC-LA	HC-SA

Thus, each manipulation (low criterion-short assignment and high criterion-long assignment) was part of a within-group, reversal design. The order between groups was counterbalanced between at least two successive course segments. That is, Group 1's high criterion-low criterion manipulation was counterbalanced by a low criterion-high criterion sequence in Group 2. Group 2's long assignment-short assignment manipulation was counterbalanced by a short assignment-long assignment sequence in Group 1.

Multiple Baseline Achievement Test

The effects of each manipulation were also evaluated using a multiple baseline testing procedure (Miller and Weaver, 1972). On the first day of class and at the completion of each of the four course parts, students completed a short-answer essay Achievement Test. The Achievement Tests were part of the course requirement in that students had to answer each question, even if the answer was trivial (*e.g.*, "I don't know"), but performance on them did *not* contribute to the student's grade. Students were informed on the first day of class that failure to complete each Achievement Test would mean that they could not progress in the course (which included not receiving a grade if they failed to finish the Achievement Test at the end of Part 4). Achievement Tests were given during class, but they were scored outside of class and no feedback was given to students about their performance on them.

The Achievement Test covered all four parts of the course and remained constant throughout the semester for all students. It consisted of 32 items—one study question item and one probe

item selected from each of the 16 content units. Thus, four study and four probe items were sampled from each part of the course. Study and probe items were similar to those described previously and both were selected to reflect major points in the course. In addition, each question consisted of two separate parts, so it was possible to award full, partial, or no credit for each answer. None of the study question items or probe items used on the Achievement Test ever appeared on a content or review quiz, but all had been used in the same course during the previous semester (Semb *et al.*, 1973), so it was possible to equate them for difficulty across the four parts of the course. Mean per cents correct for study question items used during the previous semesters for Parts 1, 2, 3, and 4, respectively, were as follows: 92.3, 94.1, 91.6, and 93.0. Performance on probe items during the previous semester averaged 62.3, 60.2, 64.0, and 61.3 for each of the four parts of the course, respectively.

Reliability

At least three times during each class day, the instructor or one of the teaching assistants selected a proctor and informed him that a reliability check would be conducted while the proctor graded a student's quiz. The observer watched the proctor as he graded each item and listened to the student's explanation of partially correct or incorrect answers. If the observer agreed with the proctor's evaluation of the student's answer, he recorded an agreement on a separate sheet of paper. If there was any discrepancy between the observer's grade and the proctor's grade, it was scored as a disagreement. Reliability checks were made on 124 content quizzes and 34 review quizzes (both groups combined). Of the 1920 items checked, there were 1742 agreements and 178 disagreements. Reliability, calculated by dividing the number of agreements by the sum of agreements plus disagreements, was 0.907.

Each item on the Achievement Test was worth 0, 1, or 2 points. To check grader reliability, a second grader scored 50 unmarked Achievement Tests from each group (10 pretests and 10 tests from each part). Agreement was defined as both graders assigning point values of 0-0, 1-1, or 2-2. Any discrepancy in points assigned by graders was defined as a disagreement. Grader reliability, calculated by dividing the number of agreements by the sum of agreements plus disagreements, was 0.925 for Group 1 and 0.928 for Group 2.

RESULTS

Percentage correct on first-attempt content and review quizzes are plotted for both groups in Figure 1. Percentage correct was calculated by dividing the total number of questions for all students combined into the number of questions answered correctly (either initially or by explanation of partially correct or incorrect answers). During Part 1 (high criterion-short assignment), Group 1 correctly answered 97.1% of the study question items and 89.8% of the probe items. When the low-criterion condition was introduced in Part 2, performance on both types of items fell to 90.2 on study items and 76.1 on probe items. Mean performance increased to 98.2% correct on study items and 94.0% on probe items when the high criterion-short assignment condition was reinstated during Part 3. During Part 4 (high criterion-long assignment), mean performance on study items was 94.0 and 79.6 on probe items.

Students in Group 2 correctly answered 91.1% of the study items and 80.3% of the probe items during Part 1 (low criterion-short assignment). Mean performance increased during the high criterion-short assignment condition (Part 2) to 97.6 on study items and 92.6 on probe items. During Part 3 (high criterion-long assignment), students correctly answered 89.1% of the study items and 81.3% of the probe items. Mean performance improved during the final high criterion-short assignment condition to 99.2 on study items and 96.7 on probe items.

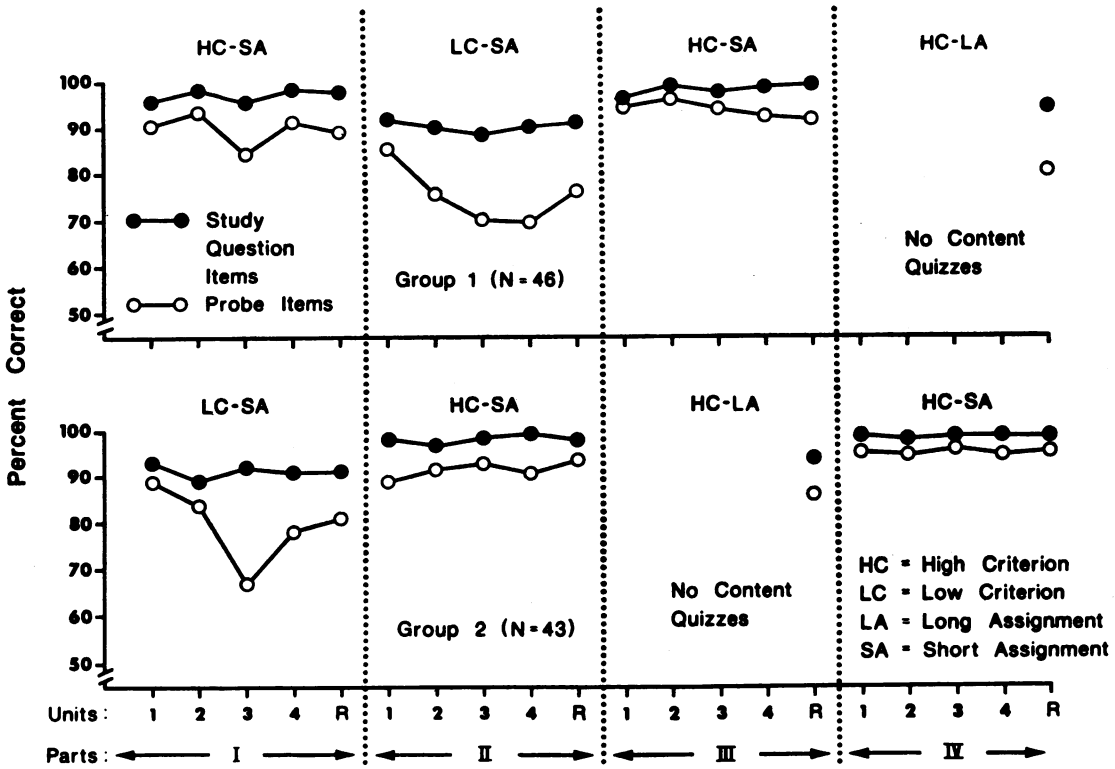

Fig. 1. Mean per cent correct for both groups on first-attempt quizzes during each of the four parts of the course. Each part consisted of four content unit quizzes (numbered) and a review quiz (R). Study question items are indicated by closed circles and probe items by open circles.

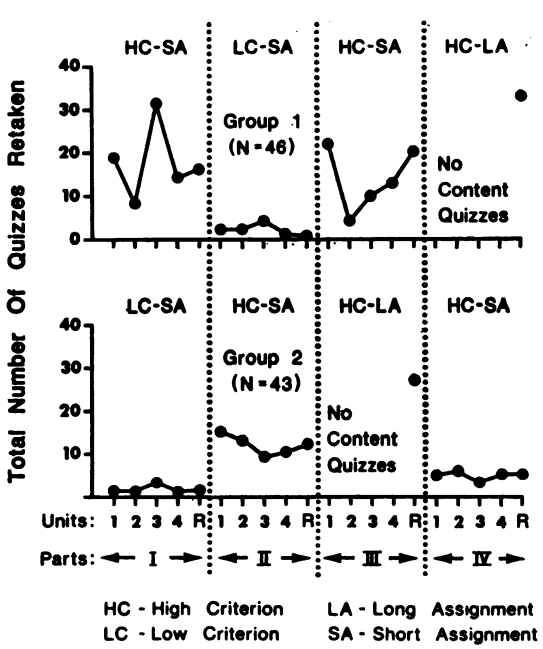

Fig. 2. Total number of quizzes retaken by both groups during the four parts of the course on content quizzes (numbered) and review quizzes (R).

The number of quiz retakes is plotted in Figure 2. The high criterion-short assignment condition produced a greater number of retakes than the low criterion-short assignment condition in both groups. Furthermore, the high criterion-long assignment condition produced a greater number of review quiz retakes than either of the other two conditions.

The results generated by the multiple baseline testing procedure are presented in Figure 3. Mean percentage correct for study and probe items is plotted for each of the five Achievement Tests for both groups. The four course parts are plotted vertically on separate ordinates, and each of the five Achievement Tests (Pretest, Parts 1, 2, 3, and 4) are shown on the abscissa. Percentages correct on both types of items were first computed for each individual student by summing the total number of points earned on each part and then dividing by the total number of points possible. The mean percentages shown

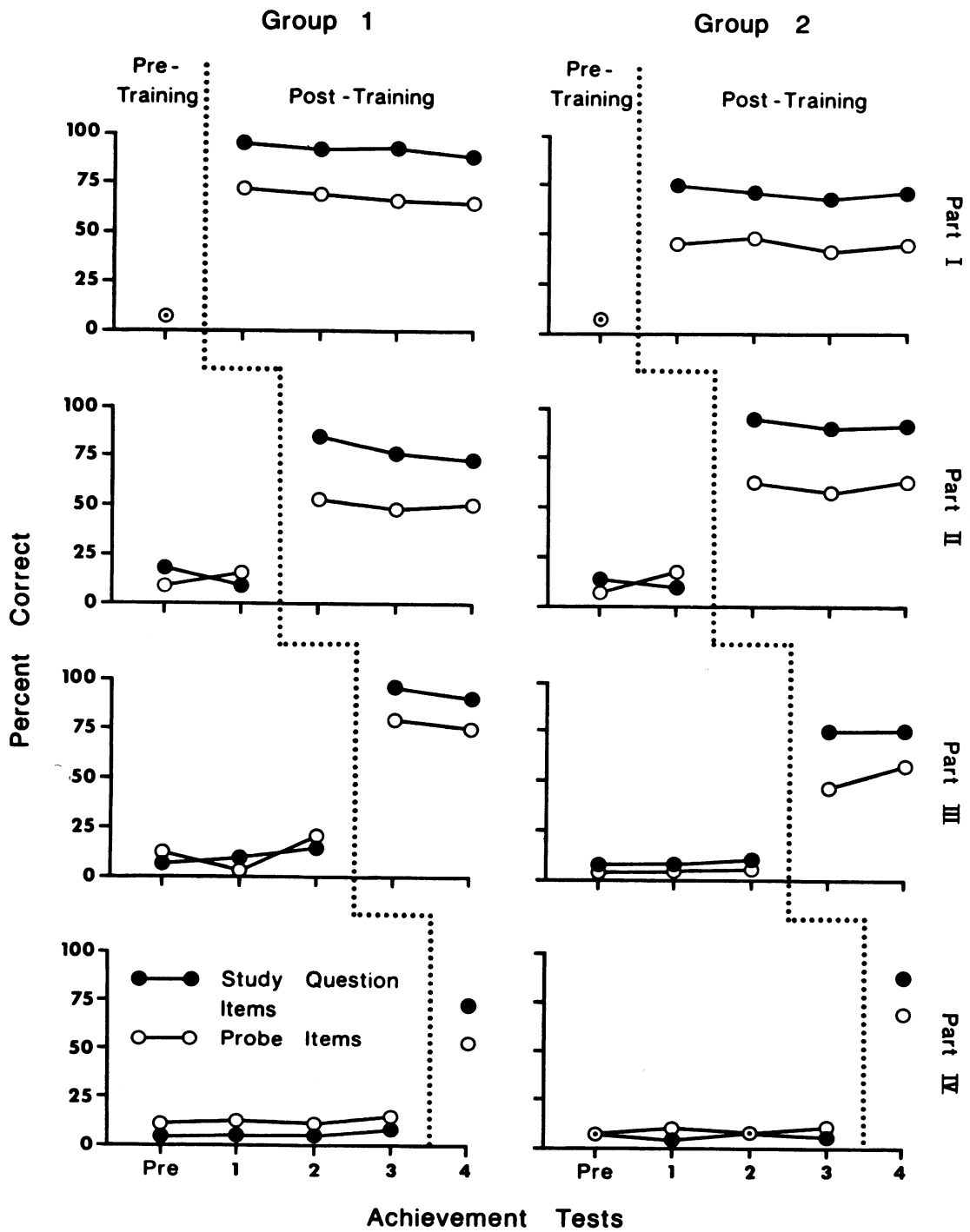

Fig. 3. Mean per cent correct on study question items (closed circles) and probe items (open circles) on each of the five Achievement Tests. Group 1 is plotted on the left and Group 2 on the right. The four course parts are plotted vertically on separate ordinates and the five Achievement Tests are shown on the abscissa.

in Figure 3 refer to the arithmetic mean across individuals.

Percentage correct on both types of items was low during pretraining conditions for each part. Increases in per cent correct were functionally related to the introduction of the instructional package for that part. For example, per cent correct on Part 2 items was low during the Pretest and on the test that followed Part 1. On the Achievement Test following Part 2, however, performance on both study and probe items for Part 2 increased markedly over their pretraining level. At the same time, performance on Part 1 items remained high, but performance on Part 3 and 4 items (ones that had not yet been trained) remained low.

During all of the posttraining conditions, both groups performed consistently better on study items than they did on probe items. Furthermore, these differences remained relatively stable on subsequent Achievement Tests throughout the semester.

Although the results from the multiple baseline testing procedure indicate that each training package produced increases in per cent correct on both types of items over pretraining levels, the effects were not the same for all training packages. To illustrate the differential effects of each experimental manipulation, percentage gains over pretest levels were calculated. Percentage gains were derived by subtracting each student's pretest performance from his Achievement Test performance on the part that had just previously been trained. Mean percentage gains were derived by summing individual student gains and dividing by N. For example, Group 1 (Figure 3) averaged 12% correct on Part 3 probe items on the pretest. After Part 3 materials had been trained, per cent correct on Part 3 materials increased to 79. The difference between 79 and 12 represents the mean gain in performance realized as a function of training on Part 3 materials. Mean percentage gains and standard deviations on both study and probe items for each of the four course parts are plotted in Figure 4.

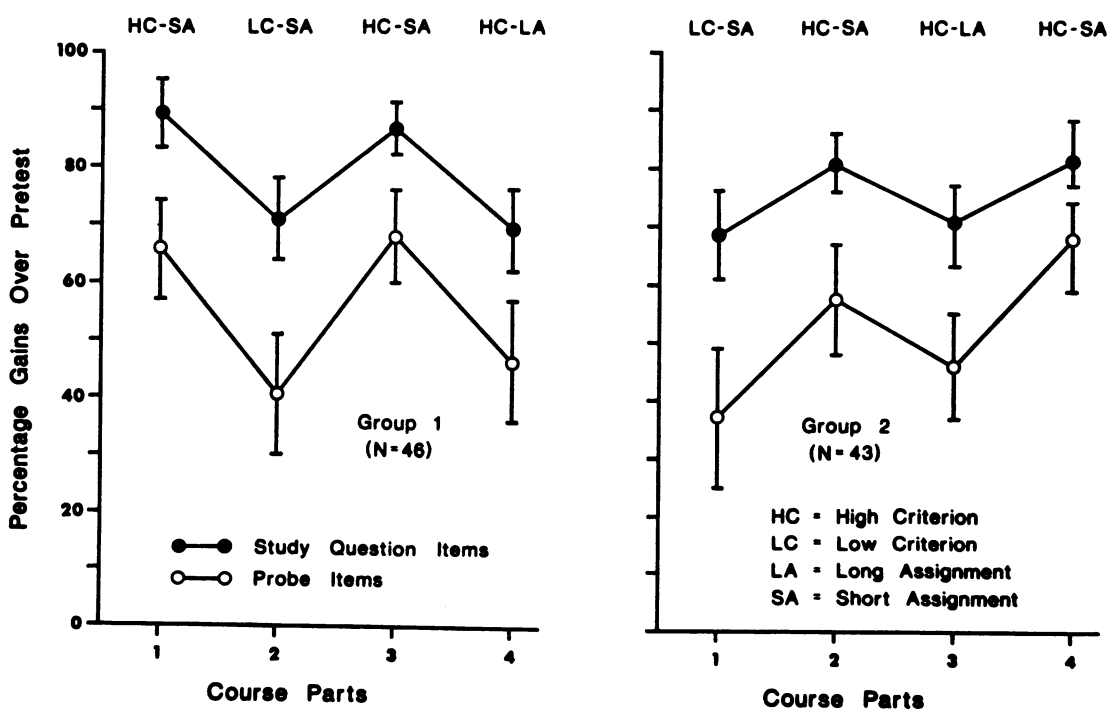

Fig. 4. Mean percentage gains over pretest levels for each of the four parts of the course. Study items are represented by closed circles and probe items by open circles. Vertical lines through each point indicate standard deviations.

During the initial high criterion-short assignment condition, Group 1 gained 88.8% on study items and 65.2% on probe items. Percentage gains on both types of items fell considerably during the low-criterion condition to 70.2% on study items and 40.6 on probe items. Gains increased during the second high criterion-short assignment condition (study = 87.2; probe = 66.8). During Part 4 (high criterion-long assignment), Group 1 gained an average of 68.9 on study items and 46.2 on probe items. In terms of individual performances, *at least* 40 of the 46 students in Group 1 changed in the same direction as the group means during Parts 2, 3, and 4 on both study and probe items.

Group 2 gained 68.2% on study items and 36.9% on probe items during the initial low-criterion condition (Part 1). When the high criterion-short assignment condition was introduced during Part 2, Group 2 gained 80.7% on study items and 57.0% on probe items. Percentage gains on both types of items decreased during the long-assignment condition during Part 3 (study = 70.3, probe = 45.4), but recovered during the final high criterion-short assignment condition (study = 81.4, probe = 67.7). In terms of individual performances, *at least* 38 of the 43 students changed in the same direction as the group means during Parts 2, 3, and 4 on both study and probe items.

DISCUSSION

The results clearly indicate the superiority of a high mastery criterion over a low mastery criterion. The low mastery criterion produced substantial and consistent decreases in percentages correct on both graded quizzes (Figure 1) and the Achievement Tests (Figure 4). Furthermore, these differences occurred on both study question (recall) items and probe items that students did not have in their possession in advance. It should be pointed out, however, that decreases in percentage gains on the Achievement Test during low mastery conditions may have resulted from the low criterion imposed on content quizzes or the low criterion imposed on review quizzes, or both. The results from content and review quizzes shown in Figure 1 suggest that the effect was a combination of both. Finally, the low mastery conditions produced far fewer quiz retakes (Figure 2) than the high mastery condition, a result that is not surprising.

The finding that a low mastery criterion (60% correct) produces inferior test performance when compared to a high criterion (100% correct) is similar to the Johnston and O'Neill (1973) study of the effects of grading criteria on academic performance. They found that students performed poorly when no criterion was imposed on performance and that increases in rates of correct responding were functionally related to increases in the value of the grading criterion. The results of both studies are similar despite differences in experimental procedures, dependent variables, and course materials. To quote Johnston and O'Neill: "the data suggest that with respect to criteria, the teacher should start high and go higher."

The long-assignment condition produced far more review quiz retakes than the short-assignment condition (Figure 2). It also produced inferior test performance on both first-attempt, graded review quizzes (Figure 2), and the Achievement Tests (Figure 4), even though the mastery criterion was set at 100% correct. It is possible that students in the short-assignment conditions learned answers to specific items on content quizzes and that those items later appeared on review quizzes; thus, they performed better on first-attempt review quizzes in the short-assignment conditions. Such an argument does not apply, however, to Achievement Test performance, because students were never exposed to those items on either content or review quizzes. Changes in item difficulty may have produced the effect, but according to performance from a previous semester's class, item difficulty was constant across the four parts of the course. Furthermore, if item difficulty was responsible, one would expect these differences

to appear in between-group comparisons over the same material, which they did not. Another possible explanation is that students who mastered the review quiz during the long-assignment conditions were not exposed to as many items from the pool, and thus did not learn as much material as when content quizzes were included. Or, it could be that long assignments produced inferior performance because proctors graded less strictly. The question is far from answered, but the present data do suggest that short assignments produce better performance than long assignments, for whatever reason.

The present results are similar to those reported by Semb *et al.* (1973). Students in both studies responded correctly to study question items far more frequently than probe items. The present data indicate that performance on both types of items can be manipulated by the kind of training the student receives.

The multiple baseline testing procedure (Miller and Weaver, 1972) used (Figure 3) proved to be effective in measuring changes in test performance as a function of each of the training procedures. The design has important implications for evaluating teaching effectiveness in a wide variety of settings. Furthermore, when percentage gains are calculated, it is possible to compare the effects of different training packages.

Finally, the present results are important validations of two components of individualized instruction (Keller, 1968), but they need not be restricted to courses taught by personalized instruction. Instructors who quiz frequently or use unit assignments may well benefit from the results of this study. In conclusion, if an instructor wishes to maximize student test performance, he or she should use a high mastery criterion and relatively short assignments.

REFERENCES

Alba, E. and Pennypacker, H. S. A multiple change score comparison of traditional and behavioral college teaching procedures. *Journal of Applied Behavior Analysis,* 1972, **5**, 121-124.

Born, D. G., Gledhill, S. M., and Davis, M. L. Examination performance in lecture-discussion and personalized instruction courses. *Journal of Applied Behavior Analysis,* 1972, **5**, 33-43.

Cramer, P. (Ed.) *Readings in developmental psychology today.* Del Mar, California: CRM Books, 1970.

Farmer, J., Lachter, G. D., Blaustein, J. J., and Cole, B. K. The role of proctoring in personalized instruction. *Journal of Applied Behavior Analysis,* 1972, **5**, 401-404.

Johnston, J. M. and O'Neill, G. The analysis of performance criteria defining course grades as a determinant of college student academic performance. *Journal of Applied Behavior Analysis,* 1973, **6**, 261-268.

Keller, F. S. "Good-bye teacher . . ." *Journal of Applied Behavior Analysis,* 1968, **1**, 79-89.

McMichael, J. S. and Corey, J. R. Contingency management in an introductory psychology course produces better learning. *Journal of Applied Behavior Analysis,* 1969, **2**, 79-83.

Miller, L. K. and Weaver, F. H. A multiple baseline achievement test. In G. Semb (Ed.), *Behavior analysis and Education—1972.* Lawrence, Kansas: Support and Development Center for Follow Through, Department of Human Development, University of Kansas, 1972. Pp. 393-399.

Miller, L. K., Weaver, F. H., and Semb, G. A procedure for maintaining student progress in a personalized university course. *Journal of Applied Behavior Analysis,* 1974, **1**.

Semb, G. *Human development lecture notes and study guide.* Unpublished monograph, Department of Human Development, University of Kansas, 1973.

Semb, G., Hopkins, B. L., and Hursh, D. E. The effects of study questions and grades on student test performance in a college course. *Journal of Applied Behavior Analysis,* 1973, **6**, 631-643.

Sheppard, W. C. and MacDermot, H. G. Design and evaluation of a programmed course in introductory psychology. *Journal of Applied Behavior Analysis,* 1970, **3**, 5-11.

Whaley, D. L. and Malott, R. W. *Elementary principles of behavior.* New York: Appleton-Century-Crofts, 1971.

Received 12 February 1973.
(Revision requested 30 April 1973.)
(Final acceptance 27 December 1973.)