*A COMPARISON OF TWO COLLEGE CLASSROOM TESTING
PROCEDURES: REQUIRED REMEDIATION VERSUS
NO REMEDIATION¹*

DARREL E. BOSTOW AND RODERICK J. O'CONNOR

UNIVERSITY OF SOUTH FLORIDA

Forty one subjects from a 10-week introductory course in Educational Psychology were randomly divided into two experimental groups. All students took weekly quizzes over content material. Members of one group received little or no academic credit if they performed at less than 90% on a weekly quiz, but could earn additional credit by taking a weekly remedial quiz. Members of the second group also took the initial weekly quizzes, but retained their raw scores and were not permitted to take the weekly remedial quizzes. Performance on a 100-item multiple-choice comprehensive final revealed a statistically significant and educationally important difference between the two groups, the required-remediation group scoring an average of one-half letter grade higher.

Several recent experimental studies have investigated the relationship between college student academic performance and various course organizations. Keller (1968) outlined a program that allowed the rate at which students completed course requirements to be determined by variables other than those controlled by the instructor ("self-pacing"). Keller's system required students to pass "readiness" quizzes or complete experiments before moving to more advanced content material. Results from Keller's program suggested that a course system with frequent quizzing and required criterion performance before proceeding to later stages in the course, produced a greater frequency of high grades than did more conventional lecture methods.

McMichael and Corey (1969) reported results that compared the "Keller System" with a more traditional lecture method involving infrequent testing, and found that the mean final

score of the Keller section was higher than the control group and that students taught by Keller's method also rated the course more favorably.

Ferster (1968) reported an "interview" instructional technique designed to develop a spoken as well as written verbal repertoire concerning elementary behavior principles. Ferster's technique incorporated students from the course in the shaping of the verbal repertoire in fellow students. All students were required to play the role of both speaker and listener for each segment of Ferster and Perrott's *Behavior Principles* (1968) text. Students were required to restudy material whenever they failed to attain adequate performance. Sheppard and MacDermot (1970) reported that a group of students experiencing Ferster's interview program scored significantly higher on final examinations than did a control group of students taught by a more conventional manner. Malott and Svinicki (1969) described a course program that incorporated many aspects of the Keller system in an elaborated format.

The present investigation involved a course design that incorporated provision for remediation when students failed to achieve adequate performance. The program of contingencies discussed was a variant of one first reported by

¹The authors wish to express their appreciation to the University of South Florida College of Education for a release time award made available for this research. We also wish to thank Dr. Jon Bailey for his helpful suggestions in the preparation of this manuscript, and R. Clifford Blair for his suggestions and participation in data analysis. Reprints may be obtained from D. E. Bostow, College of Education, University of South Florida, Tampa, Florida 33620.

Cooper and Greiner (1971) and developed by Michael (1968). While similar to the Keller and Ferster programs, in that it provided remediation, the present system did not include the "self-paced" characteristic of the Keller design. As with the Keller system, the role of the lecture in the present study was minimized in an effort to single out the effect of required remediation *versus* no opportunity for remediation on student final-exam performance.

METHOD

Subjects

Forty one students from a four-credit-hour introductory course in educational psychology at the University of South Florida were employed in the experiment during the spring term 1972. All students were non-education majors and enrolled in the course without prior knowledge that they would be involved in an experimental program.

Procedure

Subjects were randomly placed into two evaluation groups by drawing from a pile of numbered slips of paper. During the first class meeting, all students were given a 60-item multiple-choice pre-test covering the course content material. The pre-test consisted of the final exam given to a class in the previous academic term.

In the second class meeting, the first author explained that the course was going to incorporate an evaluation program that would compare student performance under two testing conditions. The class would be randomly divided into two evaluation groups and students would remain in their respective groups throughout the 10-week academic quarter. He then explained that the course content would be divided into eight full weekly units with a comprehensive final exam in the tenth week with the first unit beginning in the second week after all registered students were accounted for. All students would be required to take weekly 30-item

quizzes on the first class day of each week. Each weekly quiz would be subsequently graded and then returned during the second weekly class period. The instructor explained that the quizzes of students in Group A ("must-retake" group) would fall into any of four categories: a raw score of 27 to 30 would be assigned a Pass, 21 to 26 would be given a Questionable, 18 to 20 would earn a Rotten, below 18 would receive a Fail. A category of Pass would earn the student 10 class points, a Questionable would earn only two class points, a Rotten would earn zero points, and a Fail would cost the student one class point (−1). Students in Group A ("must-retake" group) would be allowed to take a remedial quiz on the third class meeting of any week if they failed to earn a Pass on the initial quiz of that week. The class point schedule for the remedial quiz would differ from the initial weekly quiz, in that a score of 27 to 30 raw points (a Pass) would earn only eight class points, a score of 21 to 26 (a Questionable) would earn six class points, 18 to 21 (a Rotten) four class points, and below 18 raw points (Fail) would be given zero class points. Figure 1 shows a diagram of this contingency program. The students were informed that all class points were cumulative throughout the term and that course letter grades would be based on the accumulated class point total for each student. The instructor then explained that students in Group B ("can't-retake" group) would also take the weekly initial quizzes, would return on the second class day of each week to learn their raw score on the quiz, could participate in the brief lecture-discussion with the members of Group A, but would not be required nor permitted to take the remedial quiz on the last class day of the week. Members of Group B would retain their weekly initial quiz *raw* scores, and these scores would be added cumulatively to determine final course letter grades.

The essential difference between the two evaluation groups, it was explained, was that members of Group A would receive so little course credit for less than A-level performance on the

QUIZ POINT SYSTEMS

Monday		Wednesday	Friday
Group A (Must Retake)	Initial Quiz		Remedial Quiz
	raw score class pts.		raw score class pts.
	27-30 10		27-30 8
	21-26 2		21-26 6
	18-20 0		18-20 4
	below 18 -1		below 18 0
		Return Quizzes & Optional Lecture- Discussion	NO CLASS
Group B (Can't Retake)	raw score 1-30		

Fig. 1. The weekly class activity schedule for both experimental groups and the point systems used for the accumulation of credit before the final exam.

weekly initial quiz that taking the remedial quiz was necessary to avoid the possibility of a poor final grade in the course. On the other hand, members of Group B were permitted to retain any raw score they earned on a weekly initial quiz and were not permitted to increase their point total by taking a remedial quiz.

All students were informed that the final exam would contain an approximately equal number of items from each unit of material and that it would be worth twice the value of a weekly quiz, *i.e.*, 20 class points (or 60 raw points). Consistent with the remedial cycle format to be used with Group A ("must-retake"), these students would be allowed to retake the final exam if they did not attain the 90% criterion. However, while a Pass on the first final would be worth 20 class points and a Questionable worth four points, a Pass on the retake final would be worth only eight class points, similar to the weekly class-point system.

This adjustment was contrived by the instructor in an attempt to approximate the conditions under which Group B ("can't-retake") would take the final, without appearing to sacrifice the philosophy of remediation inherent in the Group A format.

The students were again assured that each had a 50/50 chance of being placed in either evaluation group. At this point, the class was informed that those members who did not want to submit themselves to the possibility of being in either evaluation group could elect to do a number of term papers instead. All students agreed to participate in the evaluation program, and the instructor then informed each student of the evaluation group to which he belonged and passed out two forms of the course description. The basic orientation discussion was the same in both forms; however, the two forms differed in their delineation of the point systems used for accumulating credit in the course.

Table 1

The system for converting course credit into the terminal course letter grade for both experimental groups.

Group A (Must-Retake)		Group B (Can't-Retake)	
Total Accumulated Class Points	Course Letter Grade	Total Accumulated Exam Points	Course Letter Grade
90-100	A	270-300	A
80-89	B	240-269	B
70-79	C	210-239	C
60-69	D	180-209	D
Below 60	F	Below 180	F

Table 1 shows a comparison of the two methods for determining final letter grades. A course letter grade of A under both point systems represented the accumulation of 90% of the possible credit, and each successive potential letter grade represented earned credit in declining 10% intervals.

Content material for this course consisted of Whaley and Malott's *Elementary Principles of Behavior* (1971) and Malott's *Contingency Management in Education* (1971). Weekly quizzes covered between two and three chapters in *Elementary Principles of Behavior* and related parts of *Contingency Management*. Each quiz contained individual items that were either drawn directly from the questions at the end of each chapter of the texts, or were composed by the instructor during the preceding academic quarter. Approximately two thirds of the items on each weekly quiz were multiple choice, with the remaining items being short-answer essays. Remedial exams contained new items with the same proportion of multiple-choice and essay questions. The final exam contained 100 multiple-choice items, 80 of which were selected from the item pool published by Appleton-Century-Crofts to accompany the *Elementary Principles of Behavior* text. Only items that dealt with topics discussed in the summaries at the end of each chapter in the *Elementary Principles of Behavior* text were chosen for the final exam. The remaining 20 items were selected

from previous initial quizzes in the term, and dealt primarily with specific details of experiments discussed in the texts.

All quiz grading was accomplished by the instructor and two undergraduate assistants who had taken the course in the previous term. Quizzes were distributed in an unpredictable order during grading sessions with both student names and evidence of group membership being concealed when essay questions were graded. Essay questions were graded with a criterion, which was stated by the instructor before the scoring of each item. The same quiz-item number was graded by all three graders simultaneously, and each successive exam item was graded only after all graders had completed grading the preceding item. On a few occasions, the graders collectively adjusted a particular criterion when an item appeared to require an answer for which the text materials did not appear adequately to prepare the students.

A second course taught concurrently by the first author was used to obtain a rough comparison of the two weekly quizzes used in the present program. This second class was an introductory human development and learning course for 23 education majors which incorporated the same text materials. In this second class, which met 2 hr earlier than the experimental course, each week both quiz forms were distributed in an unpredictable manner to these 23 students, half of the students having one quiz form and half having the second form. Following the class, all quizzes were scored and a mean score for each of the two forms was obtained. The quiz that produced the highest mean was chosen as the weekly initial quiz for the experimental program. This procedure was contrived in an effort to preclude the possibility that remedial exams could be viewed by the students in the "must-retake" group as being less difficult.

RESULTS

At the beginning of the experiment, the "must-retake" group contained 23 students, and

Table 2
Weekly Mean Scores and Grand Means for Both Experimental Groups

Weeks								Grand Mean	
Group A									
(must-retake)									
Initial Quiz Avg.	24.22	25.80	26.30	26.80	28.40	25.00	25.62	24.62	25.84
Initial Quiz Avg. of Individuals below 27	22.94	24.33	23.25	24.20	25.50	20.90	24.30	23.17	23.57
Retake Avg.	26.70	22.53	24.12	22.08	20.66	24.27	25.46	23.10	23.62
Group B									
(can't-retake)									
Initial Quiz Avg.	25.92	25.57	26.30	26.95	26.95	23.76	25.28	25.64	25.80

Fig. 2. Frequency distributions of final-exam scores earned by members of both experimental groups. The upper figure represents data from the required-remediation group; the lower figure represents data from the no-remediation group.

the "can't-retake" group contained 24 students. However, by the third week of the quarter, six persons dropped the course, leaving 20 students in the former group and 21 in the latter. Scores made by students who dropped the course were not included in the data presented in the results. Mean scores obtained by the two groups on the 60-item multiple-choice pre-test suggested that the groups began the course with a similar verbal repertoire with respect to the content material ($\bar{X}$ for "must-retake" group = 15.41; $\bar{X}$ for "can't-retake" group = 15.62).

Table 2 presents weekly mean scores for both experimental groups. Data in this table show no consistent difference regarding mean per-

formance on initial quizzes between the two experimental groups. For comparison, the mean scores for all students in Group A ("must-retake") who earned less than 27 (a Pass) are presented. These scores, when compared with the mean scores on the subsequent retake quizzes, reveal no consistent difference in performance. The grand means are presented in the right-hand column of Table 1.

Figure 2 shows the frequency distribution of raw scores on the final exam for students in both experimental groups. The upper distribution in Figure 2 shows that final exam scores in the required-remediation group tended to be higher than those in the no-remediation group. The

Fig. 3. Scatter diagram showing the correlation between each student's entering grade-point average and his score achieved on the final exam. Regression lines have been calculated for each group's data.

mean score in the required-remediation group was 90.3, and the median was 91.5. The mean for the no-remediation group, however, was 85.9, and the median score was 86.0. The Mann-Whitney U test, when applied to these distributions yielded a Z of 2.71 ($P = <0.003$, one-tailed test).

Figure 3 presents a scatter diagram showing the correlation between each student's grade-point average upon entering the course and his score achieved on the final exam. The general scattering of scores along a diagonal from the lower-left to the upper-right hand area of the figure shows a positive correlation between the entering GPA and the corresponding final exam score for members of both experimental groups. The correlation coefficient for the "Must-Retake" group was $r = 0.47$ ($P = <0.05$). The correlation coefficient for the "Can't-Retake" group was $r = 0.62$ ($P = <0.01$).

Regression lines were calculated for both experimental groups and are presented in Figure 3. These lines, if extended, would intersect, suggesting the possibility of interaction, *i.e.*, that the required-remediation teaching method may facilitate the performance of students with lower GPAs more than those with high entering

GPAs. A statistical analysis of the difference in the angle of the two regression lines produced a $t = 0.204$, which was not statistically significant.

Figure 4 shows a comparison of students in the required-remediation group. This scatter diagram reveals a negative correlation ($r = -0.69$, $p = <0.01$) showing that the poorer a student's entering grade-point average, the more frequently he was required to take remedial quizzes. Students in this group took an average of approximately four of the possible eight retake quizzes throughout the academic quarter.

DISCUSSION

The present study suggests that a college classroom contingency system that requires students who perform poorly to retake tests over troublesome content material increases their overall achievement in the course. Inspection of the weekly performance of the two experimental groups does not reveal any consistent trend in superior performance regarding either group. Weekly quiz items were taken directly from the study questions at the end of assignments or were composed by the instructor. These weekly quiz items tended to deal with more specific details of experiments and procedures discussed in the text, whereas most items on the final were of a more general conceptual nature. The general consensus of the students appeared to be that the final exam items were less difficult. The absence of consistent differences between the initial quiz performance of students in the required-remediation group and their mean performance on each subsequent weekly remedial quiz may have been due to the system for determining which of the two quiz forms was to be the initial quiz. The choice of the more "difficult" form, as determined by administration of both forms in the second course taught by the instructor, may have obscured real increases in the performance of students who took the remedial quizzes. As no consistent

Fig. 4. Scatter diagram comparing the relationship of entering grade-point average (GPA) to the frequency of remedial quizzes required of members in Group A.

trend toward better performance on the retakes was observed, there appears to be little evidence to suggest that students were doing additional preparation for the remedial quizzes. It is possible that the required-remediation group studied more effectively for the initial quizzes in order to avoid having to show up for the remedial quiz; however, the absence of consistent differences between the two experimental groups on initial weekly quizzes suggests an insensitivity regarding the ability of weekly quizzes to reveal any experimental effect.

Two methodological problems in the present study may have influenced the results obtained. The point system used for the accumulation of course credit on weekly quizzes for the required-remediation group members involved the conversion of raw quiz scores into the Pass, Questionable, Rotten, and Fail categories, whereas the no-remediation group members simply accumulated their raw score totals. When a member of the "can't-retake" group earned a raw score of 26 on an initial weekly quiz, he kept this raw score and added it to his accumulating point total for the course. However, when an individual from the "must-retake" group earned a raw score of 26, he received a Questionable, which was worth two class points. If he then earned another 26 on the subsequent remedial quiz, he earned a second Questionable worth six class points, for a weekly total of eight class points. Eight out of 10 points is only 80% of the total possible (10), yet the raw score of 26 on either the initial or remedial quiz was actually 86.6% of the possible 30 raw points. This resulted in a slightly more stringent criterion for members of the "must-retake" group for this particular raw score value. On the other hand, a score of 27 raw points on an initial quiz earned a number of the "must-retake" group the maximum credit of 10 points, although his actual raw score (27) was actually only 90% of the total possible of 30. Thus, the conversion method used for the "must-retake" group did not precisely equate weekly quiz credit earned by both experimental groups.

While no student voiced concern about this variation in assigning weekly quiz credit, it is not known whether this factor influenced how well students in the "must-retake" group prepared for the respective weekly quizzes.

A second problem related to the opportunity for members of the "must-retake" group to retake the final exam. While the "can't-retake" group members were informed that they could not retake the final, the course description pointed out to members of the "must-retake" group that an especially high premium was placed on performance on the first final exam, as the maximum possible credit a student could earn when he found it necessary to retake the final was 60%. As the experimental groups took the final exam under these slightly different conditions, the possibility exists that the final-exam performance of either group was differentially influenced.

The apparent unusually high performance of both groups on the final exam may have been the result of several factors. The instructor made efforts to provide clear behavioral objectives that pointed the students' attention to key concepts. Also, the *Elementary Principles of Behavior* text is written in a casual but well-organized manner, which nearly all students indicated they found enjoyable. Students were informed that the extensive set of study questions found at the end of each assigned chapter would provide the foundation for many of the quiz items. It appeared from casual observation that most students used these study questions when preparing for quizzes. A brief questionnaire given at the end of the course indicated that students felt they studied more for this course than other courses, but enjoyed the class more.

The present experiment can be described as a kind of gross evaluation of remediation in college instruction. Research arising from an experimental analysis suggests that the adequacy of an organism's behavioral repertoire is directly related to the program of contingencies used to develop new behavior and to bring this

behavior under the control of appropriate variables. The remediation program described represents an educational technique that moves in the direction of respecting the rate at which students acquire specified behavior, and appears to motivate students by more careful management of reinforcement contingencies. While the dependent variable of performance on multiple-choice exam items is currently respected as a valid measure of learning in many college courses, it must be remembered that test performance is not a basic dimension of behavior. Skinner (1968) pointed out that classroom tests heavily stress response topography, while the controlling variables are often neglected. Classroom tests tend to measure the strength of memorized intra verbals, not the probability of subsequent effective practical action. Educational measurement often depends heavily on questions and answers in assessment, yet answers to questions are a relatively rare form of behavior in non-educational environments. The educator is usually more interested in influencing what practical action a student will take, given a certain set of circumstances, than he is with the development of a certain descriptive verbal repertoire about the situation. Thus, a standard classroom test situation, because it is contrived and often with aversive techniques, may give little indication of the normal probability that a student will engage in a particular behavior. Hopefully, future experimental programs will analyze more specifically the various arrangements of college instructional contin-

gencies that develop more effective practical student behavior, as well as superior classroom-test performance.

REFERENCES

- Cooper, J. L. and Greiner, J. M. Contingency management in an introductory psychology course produces better retention. *Psychological Record*, 1971, 21, 391-400.
- Ferster, C. B. Individualized instruction in a large introductory psychology college course. *Psychological Record*, 1968, 18, 521-532.
- Ferster, C. B. and Perrott, M. C. *Behavior principles*. New York: Appleton-Century-Crofts, 1968.
- Keller, F. S. "Good-bye teacher . . ." *Journal of Applied Behavior Analysis*, 1968, 1, 79-89.
- Malott, R. W. *Contingency management in education*. Kalamazoo, Michigan: Behaviordilia, 1971.
- Malott, R. W. and Svinicki, J. G. Contingency management in an introductory psychology course for one thousand students. *Psychological Record*, 1969, 19, 545-556.
- 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.
- Michael, J. L. Personal communication.
- 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.
- Skinner, B. F. *Technology of teaching*. New York: Appleton-Century-Crofts, 1968.
- Whaley, D. L. and Malott, R. W. *Elementary principles of behavior*. New York: Appleton-Century-Crofts, 1971.

Received 13 September 1972.

(Revision requested 22 November 1972.)

(Final acceptance 10 July 1973.)