8-1974

The Effects of Three Different Mastery Criterion Levels on Undergraduate Test Performance

David R. Barkmeier

The Effects of Three Different Mastery Criterion Levels on Undergraduate Test Performance

by

David R. Barkmeier

A Thesis
Submitted to the
Faculty of The Graduate College
in partial fulfillment
of the
Degree of Master of Arts

Western Michigan University
Kalamazoo, Michigan
August 1974

ACKNOWLEDGEMENTS

I wish to thank Dr. Thomas Mawhinney for allowing me to perform my study in his course and for his helpful suggestions for the implementation of this study. Special thanks go to Mrs. Susan Lipner for her grading of the test papers. I also wish to thank Dr. Jack Michael, my committee chairman, for his criticisms of many earlier versions of this manuscript and his attempt to teach me how to write.

David R. Barkmeier

MASTERS THESIS M-6120

BARKMEIER, David Raymond
THE EFFECTS OF THREE DIFFERENT MASTERY CRITERION LEVELS ON UNDERGRADUATE TEST PERFORMANCE.

Western Michigan University, M.A., 1974
Psychology, experimental

TABLE OF CONTENTS

INTRODUCTION

In 1963 a new system of instruction was developed by Fred Keller and J. G. Sherman which is called Personalized System of Instruction (PSI). Both Keller and Sherman were extremely interested in the early teaching machine research. Sherman pointed out that their mutual lack of mechanical ability kept them from seriously pursuing this endeavor (1971). However, they were considerably interested in designing new and more efficient methods of instruction without the necessity for mechanical equipment. Keller (1969) listed the basic elements of PSI as:

1. The go-at-your-own pace feature, which permits a student to move through the course at a speed commensurate with his ability and other demands on his time.

2. The unit-perfection requirement for advance, which lets the student go ahead to new material only after demonstrating mastery of that which preceded.

3. The use of lectures and demonstrations as vehicles of motivation, rather than sources of critical information.

4. The related stress upon the written word in teacher-student communication.

5. The use of proctors, which permits repeated testing, immediate scoring, almost unavoidable tutoring, and a marked enhancement of the personal-social aspect of the educational process. (P. 83.)

Keller divided the course materials into approximately 40 units, each unit being 10 to 15 pages in length. Students were required to pass a test on each unit before proceeding to the next unit. Students in Keller's system were assessed much more frequently than in traditional

courses which may assess students only once or twice a term. Thus, frequent measurement, although not specifically listed above by Keller, may be considered a sixth element in his system. Frequent measurement of the dependent variable is necessary to determine the contingencies that are controlling student performance, and may also act as an independent variable which has the effect of spreading study behavior more equally through the semester or quarter. Frequent measurement also enables an instructor to correct deficiencies in reading materials, performance criteria, behavioral objectives, or any other variables which may be contributing to poor performance before too many students are effected. Depending upon practical and other considerations these six principles are implemented in various ways and all or only a few of these principles may be incorporated into a course.

Several studies have compared the entire set of PSI components to traditionally taught classes (McMichael & Corey, 1969; Sheppard & MacDermot, 1970; Alba & Pennypacker, 1972; Cole, Martin, & Vincent, 1973). These studies demonstrated a superiority in academic performance for students participating in PSI courses. McMichael and Corey, (1969) compared one section of an introductory psychology class taught by PSI procedures with three other sections taught traditionally. Students enrolled in the PSI section scored significantly higher on a common final and on a six week follow-up test than students in the other three sections. Sheppard and MacDermot (1970) assigned students to two different sections of the same course. The final grade for the lecture or traditionally taught class was largely determined by

a final exam, while the final exam for the PSI section did not influence their final grade. The PSI group's grade was based on the number of units they completed. On the final exam the PSI group scored significantly higher than the lecture group even though the majority of the PSI students had not completed all of the units. Cole, Martin, and Vincent (1973) in a well-controlled study also showed the PSI group scoring significantly better on a common final and follow-up test than the lecture group.

In the past two years attention has turned to an analysis of the specific components of PSI which are responsible for its effects. The sixth component, unit size or frequency of measurement, was analyzed by O'Neill, Johnston, Walters, and Rasheed (1973). Three different sizes of units were systematically varied across the term. They found that the size of a unit greatly affected test performance and total study time. Test performance varied inversely and study time varied directly with unit size.

Johnston and O'Neill (1973) analyzed the second component of Keller's system, the master criterion for advancement. Keller required a perfect score on each test in order for a student to proceed to the next unit. The Johnston and O'Neill study varied three different mastery criteria or pass levels. They showed that test performance varied closely with criterion level. Students, enrolled in an abnormal psychology course, were randomly assigned to five groups. For one group criteria were not specified; subjects were told that they would be graded on the curve. For three of the groups three different mastery criterion levels were systematically presented in different

orders during the term. For the last group the criterion level for an "A" remained the same throughout the term. At equal intervals of the course lower criterion levels were introduced which corresponded to a "B" or "C". Students in this last group were allowed to proceed to further units by meeting any of the three criterion levels which corresponded to an "A", "B", or "C".

Johnston and O'Neill found that students in the first group or no-criterion group did not even match the lowest criterion specified for the other four groups. All of the students in the three groups in which the criterion level changed during the course matched or exceeded each specified criterion level. For the fixed "A" group performance decreased for some of the students after and only after the lower criterion levels were introduced.

Johnston and O'Neill made the following suggestions based on their results:

1. Define minimum criteria for academic performance.

2. Describe criteria as precisely as possible in terms of specifically defined student behaviors.

3. Define performance criteria at the beginning of the course.

4. Define criteria for only the highest standard of performance you have. If you must work in the traditional grading system, let that criteria define course grade of "A". Do not define successively lower criteria for lower course grades. (P. 268)

Davis (1973) also investigated the effects of two levels of mastery criterion on test performance with high grade point average and low grade point average students. The mastery levels were either 50% or 100%. He found that high GPA students answered correctly a

significantly higher number of questions than low GPA students, that students under the 50% mastery conditions performed significantly worse than under the 100% mastery criteria and that there was a significant GPA by treatment interaction. High GPA students scored approximately the same for the two criterion levels and low GPA students scored significantly better for the 100% criterion.

The present study is an attempt to analyze the effects of different criterion levels on test performance. It is, in a sense, a replication of the Johnston and O'Neill study but with several important differences. The procedures used in the Johnston and O'Neill study vary greatly from those used in most PSI courses. Johnston and O'Neill defined criteria in terms of rate measures in the sense of a minimum number of correct answers and a maximum number of incorrect answers per minute. While rate may be converted to percentage, rate measures are not usually used in PSI courses. The tests in the Johnston and O'Neill study consisted of fill-in-the-blank items. Most courses use a combination of different types of test items. Students in their study were tested orally by student teachers while many PSI courses use written tests. Johnston and O'Neill allowed students as many attempts as necessary to pass any one unit. In many PSI courses the number of attempts on any given unit is limited.

The present study defined criteria in terms of percentage correct, either 70, 80, or 90. Tests in the present study consisted of a combination of fill-in-the-blank, true-false, multiple-choice, and short-answer items. Answers to these tests were written, and students were allowed to make only two attempts on any given unit. Students were

not required to pass a unit before proceeding to the next unit. In essence, the present study is an attempt to replicate the Johnston and O'Neill results in a course using more normal classroom procedures.

METHOD

Subjects

The subjects (Ss) were members of a junior-level Childhood and Adolescence course conducted during fall, 1973, at Indiana University at South Bend. Thirty-two females and 22 males were enrolled in the course.

Standard Course Contingencies

The class met twice a week for 13 weeks and each class lasted approximately 75 minutes. During the first meeting of each week, the course instructor lectured on materials intended to supplement the text which was Elementary Principles of Behavior (Whaley and Malott, 1968). The reading assignment for the term consisted of the first 12 chapters of the text, one chapter, approximately 15 pages, assigned per week. The second meeting consisted of a question and answer period which was followed by a test. The test was scored either "pass", "questionable", or "fail". The "pass", the student had to answer 90% of the items correctly. If the student did so, he would then receive four "class" points towards his final course grade. A score of less than 60% resulted in a "fail" and no "class" points. Grades between 60% and 90% received a "questionable" and two "class" points. A student who failed or received a questionable was allowed to take one remedial exam which, although covering the same reading assignment, differed in detail from the "initial" exam. Percentage requirements

for a "pass", "fail", or "questionable" on the remedial were the same as requirements for the initial test. For the remedial exam a "pass", "questionable", and "fail" were awarded two "class" points, one "class" point, and no "class" points, respectively. Students proceeded to the next unit independent of having passed the preceding unit. The twelve units of readings and a term paper, worth 20 points, totaled a possible 68 "class" points. An "A" was defined as 90% of the possible points. "B", "C", and "D" were defined as 80%, 70%, and 60%, respectively. Students earning less than 60% of the possible points received an "F".

Unit Tests

Both initial and remedial tests consisted of 20 items. On each test, there was an approximately equal number of fill-in-the-blank, multiple-choice, true-false, and short-answer questions. Eighteen of the 20 test items were based on the reading material. Two questions covered the lecture. Approximately one-half of the items were taken directly from the objectives at the end of each chapter. These questions were included to meet the requirements of the course instructor. The remaining questions were constructed by the experimenter based on chapter content. Thus, students had no opportunity for prior exposure to these questions.

Experimental Course Contingencies

The experimental procedure was outlined to the class during the first meeting of the second week of the course. Students were informed

that through participating in the experiment, the required pass performance would be lower for certain units of the semester than the standard 90% correct. Students were given an option of either participating in the experiment or following standard course contingencies. All 54 enrolled students participated in the experiment.

The Ss were randomly assigned to three groups of 18. Ss were given a written statement which set forth the minimum criteria that they would have to meet or exceed in order to pass each week's test. This information was also posted on their instructor's door next to a listing of the Ss' scores for each unit.

Three levels of pass criteria were employed: 70%, 80%, and 90%. The experimental phase was in effect for nine weeks beginning the third week of class. All students were required to pass the test during the first, second, and twelfth week by meeting the standard 90% mastery criterion. A counter-balanced design was employed during the nine week experimental phase in which each group was placed under each criterion level for three weeks and then switched. This design was used to control for sequence effects and changes in difficulty of material from unit to unit. Table 1 presents the sequence followed for each group.

Dependent Variables

Dependent variables were: (1) time reported spent studying prior to each test; and (2) number of items correct on each test, with a possible range from 0-20. The lowest score received by any student was 9 correct, so the actual obtained range was 9-20.

Table 1

Mastery Criteria (%) Across Weeks

	3 - 5	6 - 8	9 - 11
Group I	70	80	90
Group II	80	90	70
Group III	90	70	80

Weeks

Students were asked to measure their own study behavior and to write the total study time per exam on their test papers. No attempt was made to differentiate between types of study or when the study behavior took place.

The tests (except for the lecture questions) were scored by a sociology major who had not previously taken the course. She was unfamiliar with the design and expectations of the experiment. The lecture questions were scored by the instructor's undergraduate assistant.

RESULTS

Four of the 54 enrolled students did not finish the course, and their data are excluded from analysis. As a result of these exclusions, Groups II and III each had 16 subjects. All remaining 50 students received an "A" for the course. Since few tests (11%) were below pass criteria, the following analysis concerns only the initial tests.

The mean for all tests taken under the 70% mastery criterion level was 17.43 (87%). For tests taken under the 80% criteria the mean was 17.69 (88%). The 90% criteria mean was 18.11 (91%). The means for the groups collapsed across criteria were for Groups I, II, and III: 18.01 (90%), 17.82 (89%) and 17.28 (87%).

A split-plot 3 by 3 analysis of variance was performed to determine if the three pass criterion levels, the sequence of presentation, and the interaction between the two had produced significantly different test performances. The null hypotheses were not rejected. All F values were less than one.

The means for the initial test of each unit are shown in Table 2. The expected results (70% criterion producing lowest scores, 90% criterion producing highest scores, and 80% in the middle) are seen for five of the nine weeks.

Of the 441 tests taken during the nine experimental weeks, 48 tests (11%) did not meet criteria. Of these 48 tests, 35 were taken under the 90% criterion, 9 were taken under the 80% criterion, and 4 were taken under the 70% criterion.

Table 2

Mean Items Correct of a Possible 20

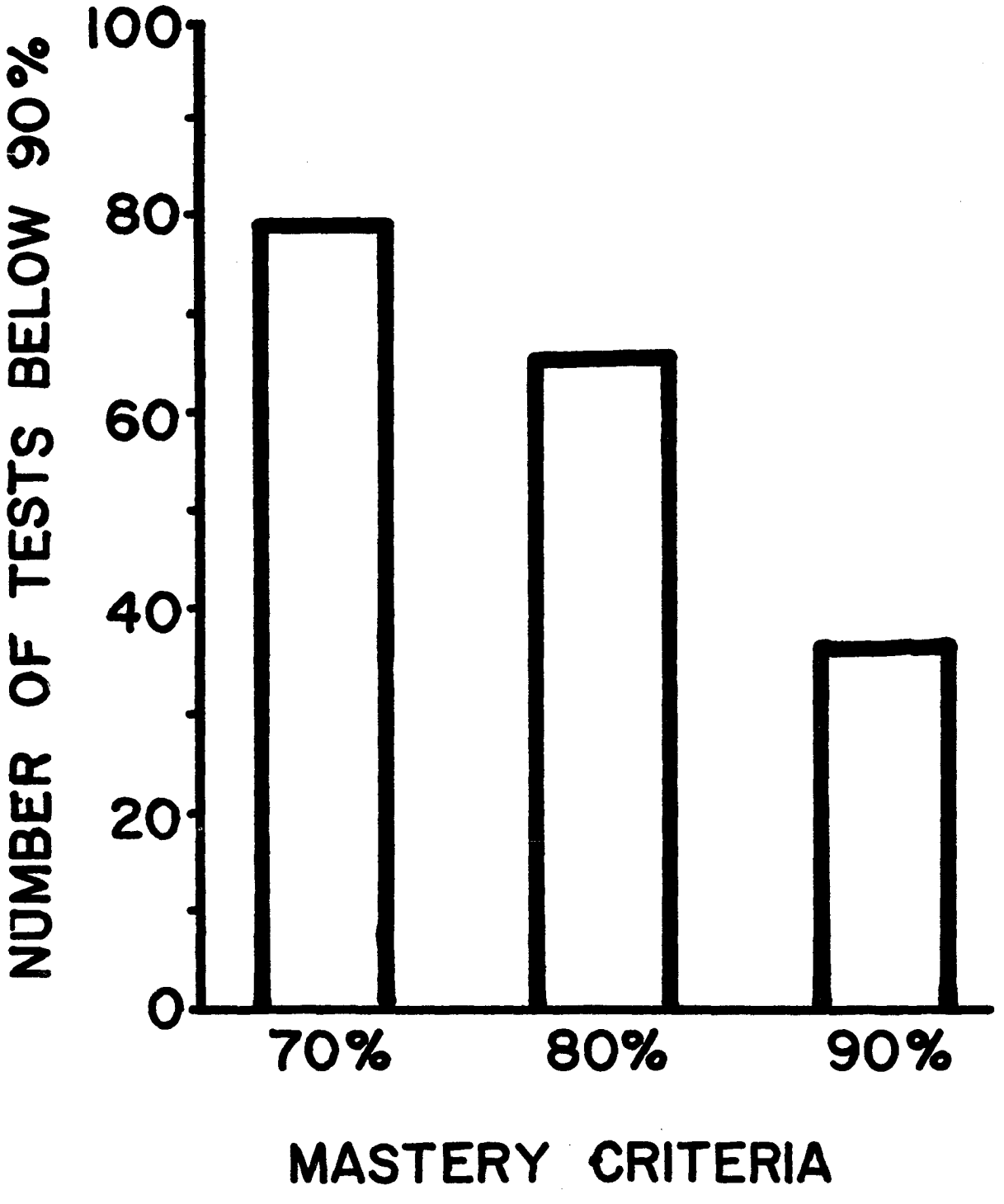
Mastery Criteria										
	70%	17.39	18.11	18.88	15.94	17.38	16.69	16.94	17.75	17.63
	80%	17.81	18.25	18.40	17.33	17.89	18.13	15.75	18.38	17.38
	90%	17.93	17.63	18.69	17.54	18.53	18.27	17.38	18.56	18.00
		1	2	3	4	5 Weeks	6	7	8	9

No significant differences were found between the means of the individual scores for each criterion level. A further test was employed to determine the effects of each criterion level on the number of tests within that level, having a score below 90%. The standard course contingency required that students score 90% correct in order to pass each unit. The number of test scores falling below 90% was obtained for each criterion level and are depicted in Figure 1. These tests would have been graded "questionable" or "fail" if the standard course contingency had been in effect. The hypothesis that there was no difference between each criterion for number of tests below 90% was tested by Cochran's Q ($Q = 21.56$, $df = 2$, $P = .001$) and was rejected.

Students answered correctly 92% of the test items which were selected from objectives at the end of each chapter. They answered correctly 88% of the test items which they had no prior exposure to. A matched - T was conducted which demonstrated a significant difference between items that they had seen and items that they had not previously seen ($T = 6.56$, $df = 49$, $P = .001$). Forty of the 50 students scored better on items that they had had prior exposure to. Of the ten students who performed in the opposite direction, only one scored more than 4% better on items that they had not previously seen than items previously seen. Students answered correctly 70% of the lecture questions. Since it is possible that the overall differences obtained between criterion levels could be due to differential responding on the lecture questions, a compilation of errors for each of the three mastery criterion levels was made. The number of errors

FIGURE 1

NUMBER OF TESTS BELOW 90% FOR EACH CRITERION



under each mastery level criterion was approximately equal.

The attempt to obtain study times was not successful. Students had been asked to write on their tests their total study time. Study times were reported on only 72% of the total number of tests taken. The mean of the study times reported was 2.11 hours per test. No attempt was made at further analysis.

Reliability checks on the true-false and multiple-choice items were conducted by the author. Reliability was calculated by the number of answers agreed upon divided by the total number of agreements and disagreements. A total of 259 (10%) answers were checked yielding a reliability of .99. Reliability assessment of the short-answer questions was complicated by the writing of a "C" on correct answers during the original grading of the tests. This necessitated that answers be rewritten before conducting reliability checks. The questions to be checked were randomly selected and rewritten from every third test. Answers were scored by a psychology graduate student who was unfamiliar with the design of the experiment. Dividing the number of agreements by the total number of agreements and disagreements yielded a reliability of .943.

DISCUSSION

Johnston and O'Neill (1973) found that criterion levels greatly influenced test performance while the present study found small differences between test performance under the three criterion levels. Tests taken under the 70% criterion level required that students correctly answer only 14 of the 20 questions on each test. The mean for all tests taken under the 70% criterion was 17.43 (87%). Tests taken under the 80% criterion were also much higher than the performance that was required. Several factors could possibly account for the difference between that which was required and the actual performance.

Unit size may be one such factor. O'Neill, Johnston, Walters and Rasheed (1973) found an inverse relationship between unit size and test performance. As unit size decreased test performance increased. The smallest unit in the O'Neill et al. study was 30 pages. Students required an average of 1.6 attempts to pass this size of unit. The unit size was approximately 15 pages in the present study. Students under the 90% mastery criterion level required an average of 1.25 attempts to pass each unit. The mean number of attempts for tests taken under all three criterion levels was 1.11.

A second factor which may have contributed to the high scores is the textbook which was used in this course. Davis (1973) used the same textbook as used in the present study in his experiment on mastery level. Students in that study were required at various points in the course to pass units by answering 50% of the test items correctly.

Students exceeded the 50% criterion requirement by an average of slightly over 30%. Johnston (1973) reported on two courses taught at Georgia State University which used almost identical procedures except for the textbook. The course which used the same textbook as the present study required an average of 1.6 attempts to pass each unit. In the other course using another textbook an average of 2.4 attempts were required to pass each unit. Another indication of the difficulty level of the textbook is the amount of study time required to meet a required level of performance. In the present study a mean of approximately two hours of study time was reported for each unit. The procedure used in the present study to measure study behavior is admittedly crude (Johnston, O'Neill, Walters and Rasheed, 1973). Even though two hours per unit is a low amount of study time, the actual time spent studying may have been less. Using the measure employed in the present study students tend to include breaks, daydreaming, or other non-study behaviors as study behavior. In addition, since the study reports were not anonymous, the students may have exaggerated study times in order to make a good impression.

A third factor which may have contributed to the high scores is that students scored significantly better on test items to which they had previous access than on items that were not made available prior to the exam. Semb (1973) reported a 10% difference in percentage correct between previously seen questions and questions not previously seen. The present study found a 4% difference. The expected difference was seen in 40 of the 50 students.

The criterion levels employed in this study is a fourth possibility for the small differences found between test performance. The rate measures used in the Johnston and O'Neill (1973) study converted to percentages are 60%, 75%, and 90%. It is possible that the criterion levels were not of sufficient difference to produce large effects.

A fifth possible factor is the amount of contact students had with the three criterion levels. In the Johnston and O'Neill (1973) study students took for the entire term an average of 21 tests. In contrast, students in the present study took an average of only 10 tests. Johnston and O'Neill demonstrated that students in the first part of the term scored much higher than the required criterion. Only during the latter part of the term were students able to predict the amount and distribution of study required to meet a specific criterion. The small number of contacts with the criteria (10) in the present study may not have been sufficient to allow the student to determine the lesser amount of study time necessary to meet the lower criteria.

The difference between criterion levels measured by the number of test scores below 90% was highly significant. Tests taken under the 70% and 80% criterion levels, though very close in actual score to the 90% criterion, would have required a re-test if the student had been under the 90% criterion. This seems to indicate that students under the 90% criterion came to the test better prepared than students under the other two criteria.

Five factors were discussed which could account for the high scores on the tests. These factors are unit size, the textbook

employed, test items which had been provided prior to the exam, the small difference between criterion levels, and the amount of contact students had with the three criterion levels.

REFERENCES

Alba, E. and Pennypacker, H. A., "A Multiple Change Score Comparison of Traditional and Behavioral College Teaching Procedures." Journal of Applied Behavior Analysis, 1972, 5, 121-124.

Cole, C., Martin, S., and Vincent, J., "A Comparison of Two Teaching Formats at the College Level." Paper presented at the First Annual Conference on Behavior Research and Technology in Higher Education, Atlanta, Georgia, 1973.

Davis, Michael L., "Mastery Test Performance and Low, GPA Student Study Time in a PSI Course." Presented at the First Annual Conference on Behavior Research and Technology in Higher Education, Atlanta, Georgia, 1973.

Johnston, James M. and Pennypacker, H. S., "A Behavioral Approach to College Teaching." American Psychologist, 1971, Vol. 26, 3, 219-244.

Johnston, J. 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.

Johnston, J., O'Neill, J., Walters, W., Rasheed, J., "The Measurement and Analysis of College Student Study Behavior: Tactics for Research." Presented at the First Annual Conference on Behavior Research and Technology in Higher Education, Atlanta, Georgia, 1973.

Keller, Fred S., "A Personal Course in Psychology." In Roger Ulrich, Thomas Stachnik, and John Mabry (Eds.), Control of Human Behavior. Glenview, Illinois: Scott, Foresman and Company, 1966.

Keller, Fred S., "Good-bye Teacher . . ." Journal of Applied Behavior Analysis, 1968, 1, 79-89.

McMichael, James S. and Corey, Jeffrey R., "Contingency Management in an Introductory Psychology Course Produces Better Learning." Journal of Applied Behavior Analysis, 1969, 3, 79-83.

O'Neill, G., Johnston, J., Walters, W. and Rasheed, J., "The Effects of Quantity of Subject Matter on College Student Academic Performance and Study Behavior." Presented at the First Annual Conference on Behavior Research and Technology in Higher Education, Atlanta, Georgia, 1973.

Sheppard, W. C. and MacDermit, H. G., "Design and Evaluation of a Programmed Course in Introductory Psychology." Journal of Applied Behavior Analysis, 1970, 3, 5-11.

Sherman, J. G., "PSI, An Historical Perspective." Paper read at the Rocky Mountain Psychological Association, 1971.

Whaley, D. and Malott, R., Elementary Principles of Behavior. New York: Appleton-Century-Crofts, 1968.