AN ABSTRACT OF THE THESIS OF

Linda Grace Thompson for the degree Doctor of Philosophy

in Science Education presented on September 24, 1975

Title: A STUDY OF THE EFFECT OF AN ANSWER-UNTIL-CORRECT MULTIPLE-CHOICE PROCEDURE ON MATHEMATICS ACHIEVEMENT

Abstract approved: Redacted for Privacy
Howard Wilson

This investigation was designed to investigate the use of an answer-until-correct procedure on multiple-choice quizzes in an independent study mathematics course. This procedure was compared to the standard multiple-choice procedure with respect to student achievement, student anxiety, and test reliability.

Using the answer-until-correct procedure students were made immediately aware of the correctness of each response they made. They continued marking alternatives on a problem until they found the correct answer. For this study, a diminishing scale of credit was assigned each question determined by the number of responses made finding the correct answer ranging from full credit on the first response to no credit on the fourth response. The tests were prepared using a chemical process which produced invisible images on the test which were brought to view by use of a special marker. When

a student marked an alternative a "+" appeared if he was correct and a "0" appeared if he was wrong.

Thirty-six students in an independent study algebra course were used in the first part of this study. During the first half of the Spring term, 1975, 15 of the students took four multiple-choice quizzes using the answer-until-correct procedure while 21 students took the same quizzes by the standard procedure. At the midterm the students in both groups took an open-ended objective examination to evaluate their achievement and responded to a five-point Likert-type scale which evaluated their anxiety toward the quizzing procedure they had used. The answer-until-correct group showed higher achievement and less anxiety than the standard group but neither difference was highly significant.

To study the comparable reliabilities of the two testing formats, 44 students in another independent study algebra class were used. A midterm examination was developed for this course which had two 15 question sections. The first section contained open-ended objective questions while the second contained multiple-choice questions which paralleled in content and difficulty the first section. When taking the multiple-choice section, 20 students used the answer-until-correct procedure, while 24 students used the standard procedure. The reliability of each method was then found by calculating the correlation of each group's multiple-choice scores with their open-ended objective scores. While each testing format was reliable the difference between the reliabilities of the two formats was not significant.

The major results of this study were:

1. The answer-until-correct procedure used on unit quizzes was slightly more effective as a teaching instrument than the standard multiple-choice procedure.
2. The answer-until-correct procedure used on unit quizzes produced slightly less anxiety toward testing than the standard multiple-choice procedure.
3. There was no significant difference in the reliabilities of multiple-choice tests when graded by an answer-until-correct or standard procedure.

Two additional results came out of this study which were not related to the hypotheses tested. First, in scoring the open-ended tests used in the study, three different scorers were asked to grade the tests independently and then the mean of these three scores was used. A high variability occurred between the three scores assigned each student. The largest deviation between the high and low score assigned a given student was 53 points on a 200 point test.

Second, in checking placement scores for this study it was found that students who elected to take intermediate algebra on an independent study basis scored significantly higher than those who elected to take it by a regular classroom basis.

A Study of the Effect of an Answer-Until-Correct Multiple-Choice Procedure on Mathematics Achievement

by

Linda Grace Thompson

A THESIS

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy

June 1976

ACKNOWLEDGMENTS

The investigator would like to acknowledge and express her gratitude to the following people:

To the students in the Mth 95S and Mth 101S courses at Oregon State University during the Spring term of 1975 who were used as subjects in this study.

To Dr. Howard L. Wilson for all the time and help given serving as her major professor.

To Dr. Gary L. Musser for his continued encouragement, advice, and great help given thoughout her graduate program.

To Dr. Norbert Hartmann for sharing his statistical expertise in the preparation of the design and the evaluation of this study.

To Kathleen Martin, Karen Swenson, Dr. Clifford Kottman, Dr. Bradford Arnold, G. Jeffrey Young, and Elsa Anderson for their assistance in carrying out the study.

To Dr. Thomas Evans, Dr. Lawrence Dale, and Dr. Ronald Cameron for their valuable services and advice given serving on her doctoral committee.

To her husband, Alan, for his patience, encouragement, and understanding during the hectic times during the study and the preparation of this dissertation.

APPROVED:

Redacted for Privacy

Associate Professor of Science Education

in charge of major

Redacted for Privacy

Chairman of Department of Science Education

Redacted for Privacy

Dean of Graduate School

Date thesis is presented September 24, 1975

Typed by Clover Redfern for Linda Grace Thompson

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure Page

A STUDY OF AN ANSWER-UNTIL-CORRECT MULTIPLE-CHOICE PROCEDURE ON MATHEMATICS ACHIEVEMENT

I. INTRODUCTION

Educators today are realizing more and more the need to consider individual differences among students when planning curriculums and programs. One highly successful instructional model which was designed to provide for these differences is the Personalized System of Instruction (P.S.I.) developed by F.S. Keller. The plan was first experimented with at the University of Brasília in Brazil in 1963 and then at Arizona State University in 1965. Today the system is being used in over 400 schools across the country in subject areas ranging from psychology to English to mathematics.

The basic elements of the P.S.I. system as Keller (1971) described them are:

> (1) the go-at-your-own pace feature which permits a student to move through a course of study at a speed commensurate with his ability and other demands upon 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; (4) the related stress upon the written word in teacher-student communication; and finally; (5) the use of proctors (student aids), which permit repeated testing, immediate scoring, almost unavoidable tutoring, and a marked enhancement of the personal-social aspect of the educational process.

At Oregon State University, the first independent study course for Intermediate Algebra II was developed in 1973 and patterned after

Keller's plan. It was administered through the newly organized Mathematical Sciences Learning Center. During the school year the course was tried, analyzed and modified. During the 1974-75 school year three other independent courses were started and administered through the center These were courses in Beginning Algebra, Intermediate Algebra I and Trigonometry. Students were given an outline of the course at the first of the term which gave course objectives with suggested sources for study. While students could move as fast as they wished they were given deadlines for completing course units. When they finished each unit they were required to take a mastery quiz. They were allowed to proceed to the next unit if they scored 70% or above. Otherwise, they were asked to restudy and take a retest.

These courses varied from P.S.I. in one aspect. Proctors, as such, were not used. Instead, quizzes were graded by the learning center clerk immediately after completion. Helpers were then available in the center to go over the quiz with the student if he wished. Students who failed a quiz were strongly urged to see a helper. These helpers were upper class mathematics students, graduate teaching assistants and professors. During any hour the center was open (48 hours per week), two or three of these helpers were on hand to answer mathematical questions for any student registered in a calculus or pre-calculus course.

It was soon obvious that an essential ingredient for these courses is a set of excellent mastery quizzes. Several alternate forms of quizzes were needed for each unit. These quizzes needed to cover all the main objectives for the unit and yet be relatively short enabling most students to finish in less than one hour. Since it was important for students to find out promptly how they had scored and a large number of quizzes were taken daily, the quizzes had to be such they they could be scored quickly and easily.

In trying to find the best type of format of quiz, short multiple-choice quizzes were tried first. While these proved to be fairly successful, there was some discontent among the student who felt it unfair to not receive partial credit on the problems. Making a slight arithmetic error meant losing full credit on a problem.

In 1974-75 a new modified form of multiple-choice was tried in the Intermediate Algebra II class. Using chemically treated paper and a special marker, students were given immediate feedback regarding the correctness of the answer selected. If they had chosen an incorrect answer, they were allowed to rework the problem and mark another response. They continued this procedure until they had selected the correct answer. A descending scale of credit was given on the problem according to the number of responses made in an attempt to choose the correct solution ranging from full credit, if the correct answer was chosen on the first response, to no credit given, if four responses were needed to chose the correct answer. (See Appendix G for sample.)

The student response to this new form of testing was extremely positive and its self scoring attribute made it very efficient to use. This study is an effort to further examine the merits of this quiz format.

Statement of Problem

This answer-until-correct (a-u-c) form of quizzing seems to have several possible benefits over regular multiple-choice quizzes. First, the immediate feedback about the correctness of the student's response suggests that these quizzes have a teaching effect much like that of programmed instruction. Not only are students reinforced when they have selected the correct answer, but they are encouraged to rework the problem, correcting errors if they have obtained an incorrect answer. This is not an attribute of regular multiple-choice (or other common used quiz formats). The first part of this study is to investigate if a teaching effect is actually present with the a-u-c format.

Second, students seem to feel more pleased and less anxious when taking quizzes in an answer-until-correct format. The frustration of selecting a wrong response because of a small arithmetic mistake seems to be lessened because they are given an opportunity to rework the problem if they get an incorrect answer. The students taking regular multiple-choice tests often complain that they lose complete credit when they know how to work a problem but make a careless mistake in the last step. A preference for the answer-until-correct format was demonstrated by the responses to a questionnaire given Winter term 1975 at O.S.U. in the independent study Mth 101S (Intermediate Algebra II). One-half of the class took standard multiple-choice quizzes and the other half took the same quizzes by the answer-until-correct method. Of the group taking standard multiple-choice quizzes 17% said they would rather have a different form of quizzes while only 3% of the answer-until-correct group said they preferred a different type. The second part of this study is to determine if the answer-until-correct format produces less anxiety in students than the standard multiple-choice format.

Third, since the answer-until-correct format gives partial credit for partial knowledge, it should be a more reliable testing instrument. The standard multiple-choice quiz often awards the same credit to the student with no knowledge as the student who works the problem by the correct procedure and makes a minor error. The answer-until-correct format, on the other hand, allows the student who makes a minor error the chance to go back, find his error, and earn partial credit on the problem. The last part of this study is to investigate if the answer-until-correct format provides more reliable scores than the standard multiple-choice format.

Definition of Terms

1. The Answer-Until-Correct Method (a-u-c) is a method of responding to multiple-choice questions where the student is made immediately aware of the correctness of his response. If he is incorrect, he is allowed to rework the problem and choose another response. In this study a diminishing scale of credit is assigned to each question determined by the number of responses made finding the correct answer ranging from full credit on the first response to no credit on the fourth response.
2. The Personalized System of Instruction (P.S.I.) is an instructional model developed by F.S. Keller which is characterized by self-pacing, mastery testing, and proctorial assistance to students.
3. Pre-Calculus Mathematics are mathematics courses which are taken prior to a standard calculus course which provide skills prerequisite to such a course. These usually include Intermediate Algebra, College Algebra, and Trigonometry.
4. The Standard Response Method for Multiple-Choice is the method of taking a multiple-choice test where the student is allowed only one response. Scored at some later time, he receives full credit for a correct response and no credit for an incorrect response.

5. Test Reliability is the consistency of measurement of a given test. In this study, a comparison for equivalence between two testing formats was made using an equivalent forms method. Using this method, the reliability of the form in question is tested by examining the correlation between scores made on it and on another test which is considered reliable.

Assumptions

1. The students which are used in this study are a representative sample of the independent study pre-calculus students who attend Oregon State University.
2. Students' response to the anxiety scale used is an accurate representation of their anxiety toward the quizzes.
3. The scores on the open-ended objective questions determined by the panel of three mathematics instructors reflects accurately a measure of students' knowledge of the subject.

Hypotheses

The following hypotheses were tested in this study:

Hypothesis 1 $H_0 1$: There is no significant difference in the overall achievement of students using a standard response method and those using an answer-until-correct method on multiple-choice quizzes during a term.

Alternate Hypothesis 1 H_a1: Students using an answer-until-correct method on a term's quizzes achieve significantly higher overall than those using a standard response method.

Hypothesis 2 H_02: There is no significant difference in the anxiety toward testing demonstrated by students responding to multiple-choice quizzes by the standard response method and those using the answer-until-correct method.

Alternate Hypothesis 2 H_a2: There is a significant difference in the anxiety toward testing demonstrated by students to multiple-choice quizzes by the standard response method and those using the answer-until-correct method.

Hypothesis 3 H_03: There is no significant difference in the reliabilities of a multiple-choice test taken by a standard response method and an answer-until-correct method.

Alternate Hypothesis 3 H_a3: A multiple-choice test taken by answer-until-correct method is significantly more reliable than one taken by a standard response method.

The first and third alternative hypotheses indicate a difference in favor of the a-u-c format because previous studies indicate this will be the case. The studies conflict on the second hypothesis, however, so the alternate hypothesis does not name the direction of difference.

Limitations

This study is limited in these ways:

1. The study is limited to students enrolled in the independent study option of Intermediate Algebra courses at Oregon State University.
2. The study is limited in time to the first half of the Spring term, 1975.

Delimitations

This study is delimited in these ways:

1. The study does not intend to investigate the use of answer-until-correct tests in areas other than pre-calculus mathematics.
2. The study does not intend to compare the answer-until-correct format to the open-ended objective test format.

Importance of Study

This study is important in that it investigates the incorporation of an effective teaching device, the immediate feedback test, into an effective teaching system, P.S.I. As will be discussed in the next chapter, the immediate feedback test has been around in various forms for over 50 years. In the studies that have been done, these tests have proved valuable teaching tools. While their use in testing with such devices as teaching machines and punchboards never really caught on, the concept is now widely used in programmed learning. One reason the immediate feedback test did not get used more widely in the past for actual classroom testing was that the devices needed were cumbersome and inconvenient to use. However, now that the new chemical process has been developed, this is no longer true.

The self-paced, independent study mathematics course seems an ideal place to make use of these tests. It carries the reinforcement principles on which this instructional model is based one step further. Students do not even have to wait to see a proctor or helper to receive feedback on their answers. In fact, by the time they see a helper they have already discovered and corrected many of their errors and can seek help only in those areas where there are genuine problems. The reworking of these problems has also forced the student to rethink the problem. A self-discovered error may make more of an impression on the student than one pointed out by another.

The mechanics of the answer-until-correct tests also make them suited to the self-study course. In these courses, quizzes are given frequently, classes can be large and immediate feedback is claimed to be more desirable. These self-scoring quizzes save valuable time for whoever ordinarily must grade the quizzes, be it proctor or clerk. It also saves proctor or helper time in that the student knows exactly what he needs to ask when he goes to the proctor.

This investigator has found no other studies concerned with this testing procedure in the setting of a mathematics independent study course. In fact, mathematics is a subject quite different in nature from the areas in which the a-u-c format has been generally investigated. At the same time, pre-calculus mathematics is being taught more and more in self-pacing programs where such a procedure could prove valuable. This study, hopefully, will give some insightful information about the effectiveness of the use of a-u-c quizzes in such programs.

II. BACKGROUND AND RELATED LITERATURE

History of Immediate Feedback Tests

The idea of tests which give immediate feedback on the correctness of answers and opportunities to make other responses is not new. In the 1920's S.L. Pressey developed a machine which would allow for just such testing. After reading a multiple-choice question the student was confronted with four numbered buttons. He would push the one marked with the number of his response. The next question would not appear until he selected the correct response but a counter kept track of the correctness of the first response to each question.

Pressey built his machine to operate in accordance with the existing knowledge about the learning process. Pressey (1927) wrote:

> . . . the 'law of recency' operates to establish the correct answer in the mind of the learner, since always the last answer chosen is the right answer. The correct response is the only response by which the learner can go on to the next question, and since whenever a wrong response is made, it must be compensated for by a further correct reaction. The 'law of exercise' is thus automatically made to function to establish the right response. . . . Finally, certain fundamental requirements of efficiency in learning are met. The learner is instantly informed as to the correctness of each response he makes (does not have to wait until his paper is corrected by the teacher).

While research such as J.C. Little's (1934) showed students utilizing the machines scored higher than did students whose papers were scored by the instructor and returned the next day, the machines never really caught on.

Instead, other methods were tried to produce the same effect. A forerunner of the chemically produced quizzes used in this study was developed by J.E. Peterson in 1930. As Peterson (1930) described his "device":

> In its simplest form this device consists of one or more sheets of multiple-choice questions whose correct and incorrect answers are treated with two matched moisture sensitive inks which turn to color when moistened. In using this device the learner reads each question, chooses from its alternative answers the one he deems correct, and touches it with a strip of moistened felt. If the chosen answer is correct, it changes immediately to a predetermined color, say red. This change of color simultaneously records his response and shows the learner whether his choice of answers was correct or incorrect.

Dr. Peterson did a study using this "self instructor and tester" using a psychology class at Kansas State College. In five comparisons made between performance with and without the self-checking feature of the tests, all comparisons showed statistically valid differences in favor of the performance with the self-checking feature. Peterson also found students expressed a strong preference for tests with the self-checking feature.

Another procedure developed by Peterson was a primitive punchboard. Students using a pin punched through an envelope and several layers of cardboard with selector holes in such a way as to record and discover their right and wrong responses.

The idea of using punchboards was further explored by Angell and Troyer (1948) beginning in 1945. They began their study after observing that the testing process of Pressey's machine was in accord with the leading theories of learning as proposed by such men as Thorndike, Guthrie, Hull and Tolman. However, they also noted that:

> . . . traditional practices of test administration will not be abandoned until simple, inexpensive, rapidly prepared foolproof devices are made available for general use.

They worked with several types of punchboard but their most successful one worked in this manner:

> This particular instrument has three parts: (a) the front cover with five perforations for each item of the test; (b) a thick cardboard middle section perforated in the same manner; and (c) the back cover which has no perforation. In order to prepare the punchboard for use, an answer sheet ($8\frac{1}{2}$" x 11") is inserted between the front cover and middle section. A 'key' is then inserted between the back cover and middle section so that red spots for the correct answer to the item will show through the proper perforations of the front cover and middle sections. . . . In this manner a student may select his answer from Item 1 and punch the corresponding spot on the punchboard. If the key shows his answer is correct, he may go to Item 2 or any other item he desires. If the selection for Item 1 is wrong, he may look over the remaining responses and continue to punch until the correct answer is located.

Angell (1949) used this punchboard experimentally in chemistry and citizenship classes at Syracuse University during the 1946-47 school year and in both cases learning was significantly enhanced by immediate knowledge of results given by the punchboard.

By 1950 Pressey (1950) has become interested in the use of punchboards for testing. His punchboard worked so that the student punched with a pencil point through a paper slip in the hole that corresponded to the answer he felt was correct. If he was right, his pencil went through the paper and down into the hole in the key sheet but if he was wrong the pencil barely broke the paper and then came up against the key sheet. Experimenting with his punchboard in Russian, English, and psychology classes, he was able to conclude,

> When the self-instructional tests were used systematically in college courses as an integral part of the teaching method, gains were substantial and sufficiently generalized to improve understanding of a topic as a whole - even help on related topics.

B.F. Skinner, who for years has investigated the shaping of human behavior, also developed a testing machine in 1958. Believing it was valuable for a student to form his own response, his machine had the student write a response and then lift a level which exposed the correct answer. His machine was based on his observations that the lapse between response and reinforcement destroyed most of its effect. With Skinner, however, the results of the former investigations were considered valuable more for the teaching aspect of the devices than for the testing aspect. Skinner's work gave rise to the multitude of programmed materials that made their appearance in the 60's and still are widely used today. The difference between these materials and the earlier work was that the testing in programmed learning materials is self-evaluation of which the teacher usually makes no use. The wise use and success of programmed instruction, however, gives strong evidence of the value of immediate reinforcement in learning.

Recent Related Studies

In the early 1970's there came a new wave of interest in using tests which gave immediate feedback for classroom testing. Two new procedures had appeared which made such tests easier to construct. One was the use of a carbon shield. Students erasing the carbon over their preferred response would find an indication of its correctness. The second method, and the one used in the study, was the use of a chemical process. A special sheet (Latent Image Transfer Sheet) used after a regular spirit master was prepared allowed one to put latent images on the tests which could only be brought to view by the use of a felt tipped pen (Latent Image Developer) containing the proper chemical.

Of the several studies which have come out in recent years, no one study investigated all three areas which are covered in this study; namely, achievement, anxiety and reliability, although each area has been studied separately. Also, none of the recent studies has been in the subject area of mathematics.

Achievement

The Navy Personnel Research and Development Center (Harding, 1974) conducted a study involving two types of quizzes for an independent, self pacing course in Aviation Mechanical Fundamentals. One group used chemically treated answer sheets which allowed the student immediate feedback on the correctness of his responses. Although no credit was given if he was wrong, he was able to mark the other alternatives until he found the correct response. When he returned his quiz, he was given a new module if he had passed, or a remediation assignment and then a retest, if he had failed.

The second group took multiple-choice quizzes where their answers were fed into the computer. They were informed whether or not they had passed after a turn-around-time of 30 to 40 minutes. If they had passed, they received no further output. If they failed, the computer gave the student a list of the general areas on which he had not met the criterion and a remediation assignment on each area and then a retest. During the turn-around-time for this group, they were allowed to continue into the next module.

At the end of the course, two comparisons were made between the two groups. The first was a comparison of the length of time taken by each group to finish the course. The second was a comparison of the scores of the final examination. The groups using the chemically treated answer sheets was found to take significantly less time finishing the course (about 15% less time). The scores of the two groups on the final showed no significant difference.

Anxiety

An investigation of students' feelings toward immediate knowledge of results on multiple choice tests was part of a study by Strang and Rust (1973). Students in an introductory course in human growth and development were asked to rate their nervousness on a ten-point scale after taking a 25 question multiple-choice quiz. Students who took the test using a carbon shield method which allowed immediate knowledge of results were significantly more nervous than those not having this knowledge. They also found that immediate knowledge resulted in loss of accuracy and an increase in the length of completion time.

The key difference, however, between this experimental setting and the a-u-c response method is that students are allowed to select other responses if they find their first response is incorrect.

Reliability

While the earlier users of immediate feedback methods allowed students to continue responding until they had found the correct response, students only received credit for their first response. The idea of allowing partial credit for responses after the first has only been experimented with recently.

One of the early studies done which attempted to give students credit for partial knowledge on a multiple-choice test was that of Frank Womer (1956). He had students respond to multiple-choice questions by selecting the distractors instead of the correct answer. Students received one point for each correct distractor selected, but received a minus three points if they selected the correct answer as a distractor. In this manner scores ranged from plus three to minus three on each question. His study, however, concerned discrimination and difficulty indices assigned to this system, so its reliability was not investigated. Also his method had no immediate feedback aspect.

Perhaps the first study testing the reliability of the a-u-c method when some credit is given for later responses was that by Gilman and Ferry (1972) of Indiana State University. A group of graduate students in education were given a multiple-choice test using a form which allowed them to continue responding until they had selected the correct response. Their tests were then graded in two ways: first, in the standard right-wrong method using only their first response to each question and second, by counting the total number of responses required to answer correctly all items on the test.

The study then compared the test reliabilities of the test using the two scoring methods. A comparison was made between odd-even item correlation coefficients and split-half reliability indices. The mean, standard deviation, odd-even item correlation coefficient and split-half reliability coefficient were all substantially larger when scored by the a-u-c method.

Gerald Hanna (1974) of Kansas State University also investigated the reliability and validity of the a-u-c procedure. Thirty-eight undergraduate students in an educational psychology class were used as subjects. Data collected included the results of (1) eleven ten-item multiple-choice quizzes (2) a 50-item multiple-choice cumulative final examination (3) two papers respectively dealing with behavioral objectives and transfer of learning and (4) an 82-item true-false interpretation exercise. Each of the two papers were subjectively evaluated by a numeric scale by the instructor. The true-false application exercise was administered without feedback and was scored objectively.

The odd-even reliability coefficient, corrected for full length, was computed for each multiple-choice measure scored by each method--standard multiple-choice and a-u-c.

To provide a validity measure, each quiz and final examination scored by each method was correlated with each of the two papers and with the application exercise. In both the case of the quizzes and the final examination the reliability and the validity was higher when scored by the a-u-c procedure although the difference in no case was significant.

It should be noted that most of the studies mentioned in this chapter were done in the area of psychology. None were done in the subject area of mathematics. One of Skinner's complaints about Pressey's machine was that multiple-choice questions too often are only written on the lower cognitive levels, those of recall and recognition. To some degree, this may be true in psychology. However, this is not necessarily true of mathematics. Most of the questions used in this study were ones where students needed to analyze the problem, apply the correct procedure, work the problem and then look for a correct response. Looking at the answers before doing this usually gave no clue to the test taker as to the right answer. The questions still demanded a written response on the part of the student before answering, something Skinner was convinced was very necessary for optimal teaching by the instrument.

It should also be noted that the Navy Personnel Research and Development (1973) study was the only other study where quizzes were used in an independent study setting similar to the one of this study. Since this study is particularly concerned with the use of a-u-c quizzes in independent study programs, it is important that they be investigated in such an environment.

Summary

The immediate feedback test which has developed into the a-u-c quizzes in this study was developed in accordance to the following theories of learning:

1. Learning is facilitated by immediate reinforcement to a correct response. (Upon marking a response, the appearance of the latent image informs the student immediately of the correctness of his answer.)
2. That which is encountered most recently is remembered best. (Since a student continues to mark responses until he finds the correct one, the last response he marks is always the right one.)

The following conclusions can be drawn from the literature:

1. In the majority of the studies, use of quizzes which gave immediate feedback of results improved student achievement.
2. While one study reported higher student anxiety on immediate feedback quizzes, other studies have reported student's preference for them.
3. Tests graded by the a-u-c method had a higher reliability and validity than those scored by the standard method.

Most of the studies differed from the study of this dissertation in the following ways:

1. No other study used tests in the subject area of mathematics. As noted, mathematics lends itself to a different nature of questions than the areas of psychology and education which were most commonly used.
2. Only one other study dealt with a-u-c quizzes in the setting of an independent study course.
3. Only two other studies granted students any credit for later responses on a-u-c quizzes.

III. THE STUDY

This chapter will be divided into three main sections. Each section will give the design of the study used to test one of the hypotheses. The first will be the study on achievement, the second on anxiety and the third on reliability.

Study on Achievement

The Population

For this study the students in Mth 101S, the independent study section of Intermediate Algebra II, at Oregon State University during the Spring term 1975 were used. Of the 40 students who began the study, 36 students completed it. The other four were students who dropped the course early in the term.

The Design

The research design for this part of the study is a posttest only, control group design illustrated by

$$\begin{array}{ccc} R & X & O \\ R & & O \end{array}$$

This indicates the use of two groups randomly selected (R). One is the experimental group which will receive the experimental treatment (X). The second is the control group which will receive the same treatment with the exception that the experimental treatment is withheld. After the experimental treatment is given an identical posttest (O) measuring the aspect of concern is given both groups.

For this study the students in the population were divided into two groups by use of a random number table. During the first half of the term, the students in the first group were given four ten-question quizzes using the answer-until-correct method.[1] There were 17 students in this group of which 15 completed the study. The students in the second group were given the same quizzes using the standard multiple-choice method. There were 23 students in this group of which 21 finished the study.

On the quizzes this question format was used: (see Appendix G)

1. 5! =

___A. 15 ___ B. 120 ___ C. 20 ___ D. 60

Students using the answer-until-correct technique would use the Latent Image Developers to answer the question marking the blank in front of the answer they decided was correct. Latent images had been placed on each blank of the question and a "+" would appear if the

[1] The a-u-c quizzes were prepared using Latent Image Transfer Sheets produced by the A B Dick Company. Using the sheets after a spirit master has been prepared enables one to add latent images to the quizzes. These images are invisible and can only be brought to view by use of a Latent Image Developer, a felt tipped marker containing a special chemical which reacts with the chemical put on the quizzes by the transfer sheets.

student was correct while a "0" would appear if he was incorrect. He would continue marking responses to any given question until he found the "+".

Students using a standard multiple-choice took the very same quizzes only they were not given Latent Image Developers. Instead, they put a mark in pencil in front of the response they felt was correct. They only made one response to each question and were given no feedback about the correctness of their responses until they had finished the entire quiz.

At the middle of the term both groups were given the same midterm examination over the material which had been covered on the four quizzes. The examination was a 20 question open-ended objective test. A jury of four members of the mathematics faculty at O.S.U. who had a knowledge of this course and its content examined the test to verify its validity concerning material covered in the first half of the course. They also made suggestions for the improvement of the test. (These faculty members were Clifford Kottman, Bradford Arnold, Elsa Anderson, and Gary Musser. The final test is found in Appendix A.)

After the tests were taken they were graded independently by three mathematics faculty members (Clifford Kottman, Bradford Arnold, Karen Swenson). They were asked to grade independently the tests at a single sitting and strive for as much consistency as possible in their grading. They made no marks on the tests themselves. They each were given a class roster on which to record their scores. The three scores for each student were averaged and this mean was considered the student's score on the open-ended objective examination. This procedure, namely that of averaging the scores of the jury members, was followed since there is some indication that there can be considerable variability in scoring open-ended mathematics test items. The wish was to get as accurate a picture of the student's knowledge as possible.

The means and variances of the two groups were then calculated and compared. Because a comparison of means was desired a standard t-test was used. Since the alternate hypothesis expressed the direction of difference, a one-tailed test was used.

Generalizability

Since the results of this study were to be generalized to the total population of independent study pre-calculus students at Oregon State University, a study was made of the initial placement scores of the group used in this study as compared with the total population in Mth 101S during the 1974-75 school year. A chi-square test was used to test if this sample group could be considered representative of the total population at least as far as their initial placement test was concerned.

Study on Anxiety

The Population

The population in this section of the study was also the student enrolled in Mth 101S, Intermediate Algebra II at Oregon State University during Spring term 1975 The study was begun with a group of 40 but ended up with 36 at the midterm when the study was concluded.

The Design

The design for this study is also a posttest-only, control group which can be illustrated as:

R X O

R   O

The same groups as were designated for the study of achievement were used in this study and the experimental treatment was the same, namely, the use of the answer-until-correct procedure on four quizzes during the first half of the term.

The posttest used here was a brief questionnaire on student anxiety toward testing. Each student responded to a simple Likert-type scale where students rated their anxiety toward the quizzes they had taken on a scale from one to five. A precedent for this type of questioning was the study of Strang and Rust (1973) who studied the effect of immediate feedback on test takers by using a ten-point scale for anxiety. Students were asked to respond to this scale before they took their midterm examination. The means and variances for the two groups were calculated. The means were then compared and the difference was tested for significance by a standard t-test. Since the direction of difference was not specified in the alternate hypothesis, a two-tailed test was used.

Study on Reliability

The Population

The population for the study was the students enrolled in Mth 95S, the independent study section of Intermediate Algebra I at Oregon State University during the Spring term, 1975. The students used were the 44 students enrolled in the course at the time of the midterm examination.

The Design

To determine the relative reliabilities of the two testing methods, answer-until-correct and standard multiple-choice, scores received on tests using these formats were compared to those received on comparable open-ended objective tests.

This was accomplished on the midterm examination for the Mth 95S class. A two part midterm examination was written. One part contained 15 open-ended objective questions covering the content of the first half of the Intermediate Algebra I course. The other part contained 15 multiple-choice questions, each question paralleling in content and difficulty one of the open-ended questions. Five mathematics faculty members who were familiar with the course content juried the test to verify that it validly covered the content of the first half of the course and that the two parts were truly parallel in nature. They also made various suggestions for the improvement of the test. (These instructors were Karen Swenson, G. Jeffrey Young, Clifford Kottman, Gary Musser, and Howard Wilson.)

The multiple-choice portion of the examination was piloted in a second-year algebra class at a local high school.[2] An item analysis was then made on the basis of the results of this test to investigate the difficulty and discrimination of the test items. An item was considered to have good discrimination if its discrimination index was greater than 40 and good difficulty if it had a difficulty index over 80. All items on the test proved to have either high difficulty or discrimination. (The item analysis is in Appendix D.)

The intent during the Mth 95S midterm was to have all students take the open-ended part of the test and then have one-half of the students take the multiple-choice section by a standard method and

[2] Thirty-eight students in the class of Paul Ashdown at Corvallis High School took the examination.

one-half take it by the answer-until-correct method. After scoring, correlations would then be computed between the two sections of the test for each group to determine which method of taking the multiple-choice section had the higher correlation with the open-ended section.

Since it was possible that students might receive help on one section of the test by the questions on the other section, the students took one section at a time and turned it in before receiving the other. To test if such teaching interaction was occurring, one-half of each group took the multiple-choice section first and the other half took the open-ended section first.

Essentially then there were four groups in this part of the study:

Group 1: Students taking the answer-until-correct section first, then the open-ended section

Group 2: Students taking open-ended first, then the answer-until-correct section

Group 3: Students taking the standard multiple-choice section first, then the open-ended section

Group 4: Students taking the open-ended section first, then the standard multiple-choice.

Since the four groups were chosen by using a random number table, there was variety in the size of the groups ranging from 9 to 13.

After the test was administered, all the open-ended sections from both groups were graded independently by three mathematics faculty members (Clifford Kottman, Karen Swenson, Bradford Arnold). As in the achievement study, they were urged to strive for consistency and to grade all tests at one sitting. Again their three scores were averaged and this mean was considered the student's score on the first section. Each problem was scored on a ten-point basis so this section had a maximum score of 150.

The multiple-choice section was also graded on a ten-point per question basis. Students who took it by the standard method received ten points if they had a question correct and no points if they missed it. Students who took the multiple-choice section by the answer-until-correct method received ten point if they got a question correct on the first response, five points if they got it correct on the second response, two points if they got it correct on the third response, and no points if they got it correct on the fourth response

A 2 x 2 factorial analysis of variance was run on the four groups' mean scores on the open-ended section to see if there was a significant teaching interaction caused by which section of the test was taken first.

Groups 1 and 2 were then combined into a total a-u-c group and Groups 3 and 4 were grouped into a total standard multiple-choice group. Correlations were calculated between the scores on the open-ended section and the multiple-choice section for each group. A comparison for significant difference was made by conversion of these correlations into z-scores and the use of the Standard Normal Distribution.

A second test of the reliability hypothesis was made by rescoring the multiple-choice sections of Groups 1 and 2 by a standard multiple-choice procedure. The correlation coefficient between these scores and the open-ended section scores was then calculated. This correlation was then compared to the correlation previously calculated between the a-u-c scores on the multiple-choice section and the open-ended section. The comparison was made by transforming these correlations to z-scores and using the Standard Normal Distribution.

IV. FINDINGS

This study was undertaken to investigate aspects of the answer-until-correct testing format. It compared the a-u-c format to the standard multiple-choice format with respect to anxiety production, reliability, and teaching ability. The three hypotheses tested were:

$H_0 1$: There is no significant difference in the overall achievement of students using a standard response technique and those using an answer-until-correct technique on multiple-choice quizzes during a term.

$H_0 2$: There is no significant difference in the anxiety toward testing demonstrated by students responding to multiple-choice quizzes by the standard response method and those using the answer-until-correct method.

$H_0 3$: There is no significant difference in the reliabilities of a multiple-choice test taken by a standard response method and an answer-until-correct method.

This chapter will be divided into four sections. The first three will discuss findings concerning the hypotheses. The fourth section will discuss findings unrelated to the hypotheses.

Achievement Study

Generalizability

The first concern in this section was that the sample used, the Mth 101S class Spring term 1975, was representative of the independent study pre-calculus population at Oregon State University. Therefore, the scores on a placement test given to all Mth 101S students at the beginning of each term were compared. The scores for the sample group were compared to those of the total population in Mth 101S during the 1974-75 school year. A chi square (χ^2) goodness of fit test was used.

The possible scores on the test of 15 questions were divided into five groups. Using the large population as the normal, the expected percentage of the total falling in each group was calculated. The test criterion was calculated by

$$\chi^2 = \Sigma(f_i - F_i)^2 / F_i$$

where f_i represents the number of scores in the ith group in the sample and F_i is the number predicted by using the percentages calculated from the large population.

Table 1. Comparison of sample to total population.

	Number of Answers Correct					
	0, 1, 2	3, 4, 5	6, 7, 8	9, 10, 11	12, 13, 14, 15	Total
Total population	3	21	52	32	15	123
Percent	2.4%	17.1%	42.3%	26.0%	12.2%	
Calculated F_i	.83	5.80	14.34	8.85	4.15	
Sample f_i	1	4	10	13	6	34

$$\chi^2 = 4.75 \quad \text{D.F.} = 4 \quad P < .67$$

Therefore, based on student scores on the placement test, it can not be rejected that these students were a representative sample of the total Mth 101S population.

Reliability Study

To compare students taking a-u-c quizzes and standard multiple-choice quizzes with respect to overall achievement, an open-ended objective examination was given. The content validity of this test was checked by having a jury of four O.S.U. mathematics instructors examine it. The reliability of the test was judged by calculating the correlation coefficient between students' scores on the odd items and even items. For this calculation the scores assigned by

Dr. Clifford Kottman[3] were used.

The correlation coefficient was calculated using the formula

$$r = \frac{\Sigma x_{i1} x_{i2}}{\sqrt{\Sigma x_{i1}^2 \Sigma x_{i2}^2}}$$

where x_{i1} and x_{i2} are the ith student's score on the odd and even problems respectively. The resulting r was .76. The Spearman-Brown formula was then applied to find the reliability on the full test.

$$\text{Reliability on full test} = \frac{2 \times \text{Reliability of } 1/2 \text{ test}}{1 + \text{Reliability of } 1/2 \text{ test}}$$

$$= .86$$

Since teacher-made tests commonly have reliabilities between .60 and .85, this test was considered reliable.

Achievement Results

The midterm examinations were scored independently by three instructors and these grades were then averaged and the means assigned as the students' scores. The scores of the group who had previously taken a-u-c quizzes were then compared to the scores of

[1] Due to an oversight in directions given to the scorers, Dr. Kottman was the only scorer who kept a record of individual deductions on problems.

those who had previously taken standard multiple-choice quizzes. The results are below in Table 2.

Table 2. A-U-C and standard groups midterm achievement.

Group	Mean on Midterm	Variance on Midterm	Number in Group
a-u-c	137	1051	15
Standard	123.5	2075	21

t = .98 D.F. = 34

This t-value is significant at the .14 level. Therefore, while the a-u-c group performed better on the midterm, the difference in results was not highly significant.

Anxiety Study

Since the same groups were used in this study as the previous one, the comments in the last section dealing with the generalizability of the study apply to this study as well.

The anxiety students felt toward quizzes was measured by the students' response to a five-point Lickert-type scale where five represented the highest anxiety level. The results of the study are given in Table 3.

Table 3. A-U-C and standard anxiety.

Group	Mean	Variance	Number in Group
a-u-c	2.53	1.23	15
Standard	3.00	1.58	21

t = 1.15 D.F. = 34

Using a two tailed t-test, 1.15 is significant at a .26 level. Therefore, while the a-u-c groups showed less anxiety the difference between the two groups was not highly significant.

Reliability Study

Reliability of Open-Ended Test

Since in this study a comparison of the results of the two types of multiple-choice examinations to the results on the open-ended was to be made, it was necessary to confirm that the open-ended test itself was a valid reliable test. The content validity was confirmed by a jury of four mathematics instructors. Its reliability was checked by calculating the correlation coefficient between the scores made on the odd and even problems of the test.

The resulting r was .56. Using the Spearman-Brown formula the reliability on the full test was .72. Since teacher-made tests commonly have reliabilities between .60 and .85, this test was considered reliable.

Comparison of A-U-C and Multiple-Choice Reliability

Four groups were used for this study. Each group took the two part midterm examination. One part was open-ended objective questions. The second part was multiple-choice questions which paralleled in content the questions on the first part. The groups varied in the order in which they took the two parts and in the procedure used taking the multiple-choice part. The four groups were:

Group 1 - Multiple-choice (a-u-c) taken first, open-ended last

Group 2 - Open-ended taken first, multiple-choice (a-u-c) last

Group 3 - Multiple-choice (standard) taken first, open-ended last

Group 4 - Open-ended taken first, multiple-choice (standard) last.

The resulting mean scores on the sections and the correlation coefficients between the two parts are given in Table 4.

Table 4. Correlations between multiple-choice and open-ended.

Group	Number	Mean on Open-Ended	Mean on Multiple-Choice	Correlation Between Parts	Pooled Correlations
1	9	116.2	127.3	.4371	.6510
2	11	110.1	124.7	.8285	
3	9	95.0	100.0	.9403	.7569
4	15	97.2	107.3	.6241	

The scattergrams of the four groups are shown below.

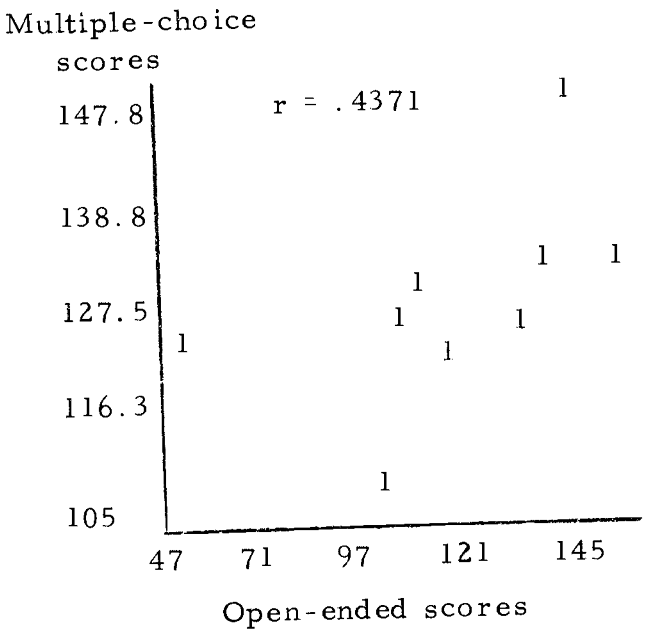


Figure 1. Group 1 scores.

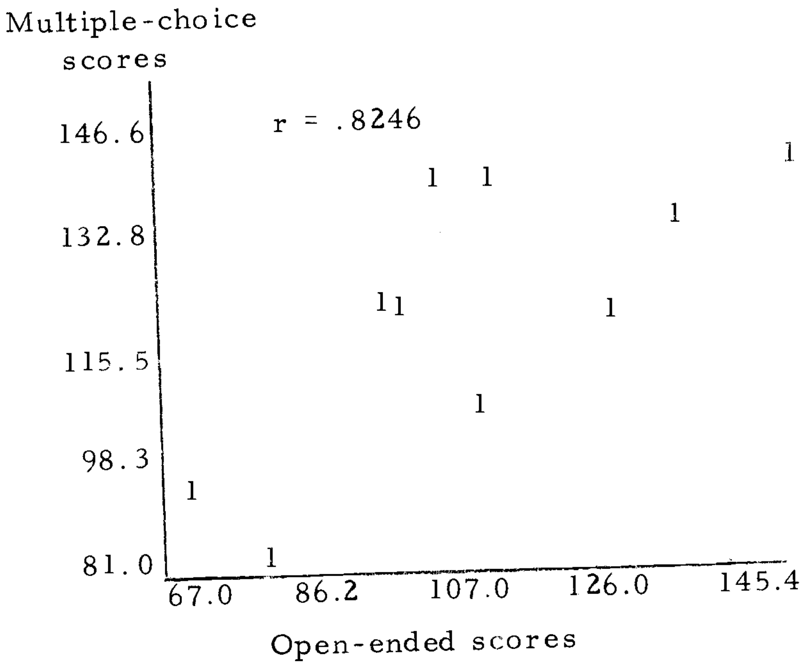


Figure 2. Group 2 scores.

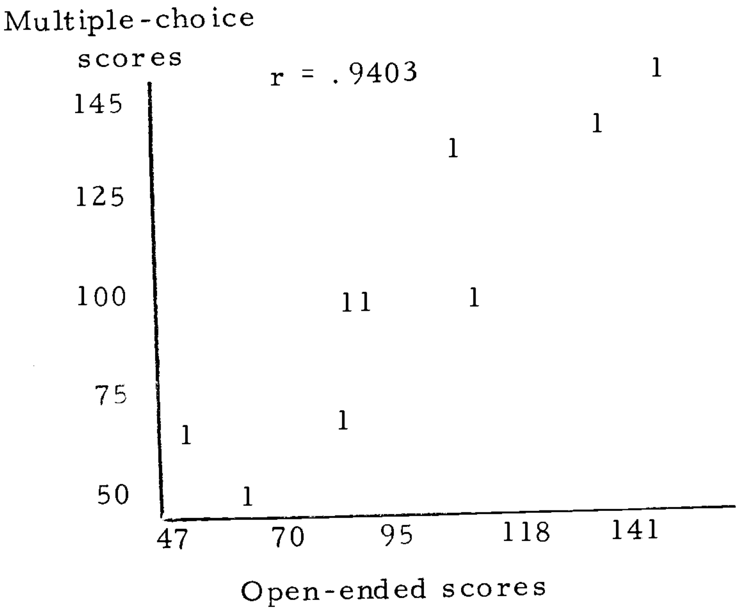


Figure 3. Group 3 scores.

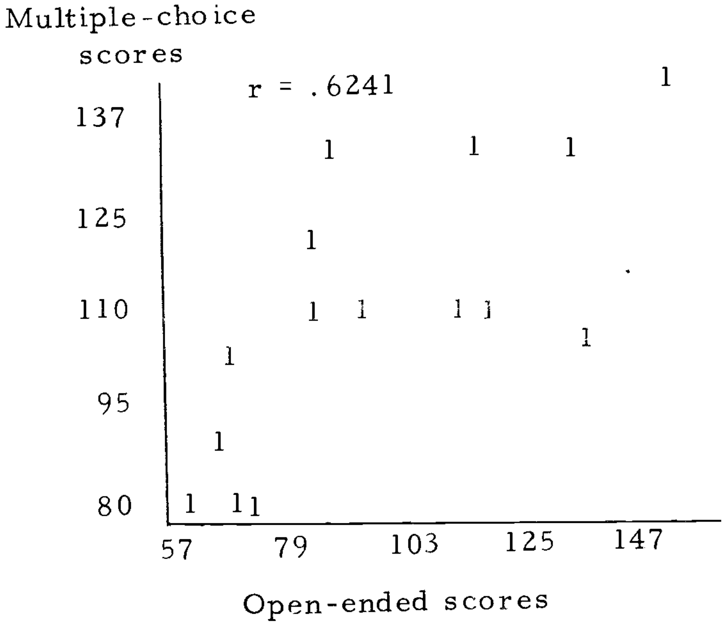


Figure 4. Group 4 scores.

The low correlation on Group 1 is explained partially by a student who scored 47 on the first part and 122 on the second since all the rest of the points fall approximately along a line.[4] A truer picture of the situation may then be derived by removing this anomaly. An adjusted table appears below.

Table 5. Adjusted table of correlations.

Group	Number	Mean on Open-Ended	Mean on Multiple-Choice	Correlation Between Parts	Pooled Correlations
1	8	124.9	128.0	.6225	.7657
2	11	110.1	124.7	.8285	
3	9	95.0	100.0	.9403	.7569
4	15	97.2	107.3	.6241	

The difference in the means on the open-ended section suggested there was a teaching effect present when students took the answer-until-correct section first. Using the results on the open-ended section a 2 x 2 factorial analysis was made to test for a possible interaction.

Table 6. Means on open-ended section.

Method Used on Multiple-Choice	Part of Test Taken First			
	N	Multiple-Choice	N	Open-Ended
a-u-c	8	124.9	11	110.1
Standard	9	95.0	15	97.2

[4] The 47 score was consistent with this student's other work in the course, so the high score on the a-u-c section can probably be attributed to chance.

The within-classes mean square, computed from the individual observations was 742.6 with 39 d.f. To test the interaction, it was computed from the sub-class means in the usual way for a 2×2 factorial:

$$124.9 + 97.2 - 110.1 - 95.0 = 17.0$$

Taking account of the sub-class numbers, the standard error of this estimate is

$$\sqrt{s^2(1/8+1/9+1/11+1/15)} = 17.1$$

The value of t is $\frac{17.0}{17.1} = .994$ with d.f. 39, P about .34. We, therefore conclude the interaction is not significant and continue our study.

Since the interaction was insignificant we can combine Groups 1 and 2 to consider the correlation between open-ended and a-u-c tests. This correlation as noted in Table 5 is .7657. Also, we can combine Groups 3 and 4 to consider the correlation between open-ended and standard multiple-choice tests. This correlation was .7569.

To compare these two correlations for significant difference we convert them first to z-scores by the transformation

$$z = \frac{1}{2} \log_e (1+r)/(1-r)$$

The test for significance between these z-scores is given below.

Table 7. Test of difference between correlations.

Group	Number	r	z	1/(n-3)
a-u-c	19	.7657	1.01	.0625
Standard	24	.7569	.989	.0417
		Difference	.021	.1042 Sum

$$\sigma_{z_1 - z_2} = \sqrt{.1042} = .32 \quad z = \frac{.021}{.32} = .07 \quad P = .53$$

Therefore, there is no significant difference between the two. It should be noted, however, that the correlations from both groups are significant at the .01 level. Therefore both tests were reliable tests.

One further test of the third hypothesis was made. The answer-until-correct tests of Groups 1 and 2 were regraded by a standard multiple choice method. In other words, they were regraded only counting students' first choices. Table 8 exhibits these results.

Table 8. Correlations of regraded Groups 1 and 2.

Group	Mean Graded by a-u-c	Mean Graded by Standard	Correlation a-u-c vs. Open-Ended	Correlation Standard vs. Open-Ended
1	128.0	112.5	.6225	.5900
2	124.7	111.8	.8285	.7597
1 & 2	126.1	112.1	.7657	.6774

The test for significance between the correlations for the total group (Groups 1 and 2) is given in Table 9 below.

Table 9. Test of difference a-u-c and regraded standard.

Groups 1 & 2 Graded by	r	z	n	1/(n-3)
a-u-c	.7657	1.01	19	.0625
Standard	.6774	.82	10	.0625

Difference .19 .1250 Sum

$$\sigma_{z_1 - z_2} = \sqrt{.1250} = .354 \qquad z = \frac{.19}{.35} = .54 \qquad P = .21$$

While the difference in correlations here were more pronounced in favor of the a-u-c method, it is still not highly significant.

Findings Not Related to Hypotheses

Independent Study Versus Standard Classroom

In examining the results of the placement test for the generalizability of the study, it was noted that students electing to take the Mth 101 course by independent study had a higher mean score on the placement test than those who took it by a standard classroom procedure during the 1974-75 school year. The table below shows this difference is significant at the .05 level.

Table 10. Classroom vs. independent study on placement test.

Group	Number	Mean on Placement	Variance
Standard class	572	7.371	8.627
Independent study	113	7.976	8.691

$t = 2.03$ d.f. $= 683$ $P < .05$

Variability of Grading Open-Ended Tests

A second finding of this study unrelated to the hypotheses was that there occurred large variability in the scoring of the open-ended objective questions. The three scorers, two professors and a graduate teaching assistant, were requested to grade each question on each examination on a ten-point basis. On the 150 point Mth 95 open-ended test the mean deviation between the highest and lowest score assigned a given student's test was 15.3 points and the largest deviation between the three scores was 52 points. On the 200 point Mth 101 open-ended test the mean deviation between the highest and lowest score given each test was 22.36 and the largest deviation was 53 points. On both test deviations in scoring averaged over 10% of the total score.

Summary of Results

1. While students taking a-u-c quizzes during the term scored higher on the open-ended objective midterm on the average than students taking standard multiple-choice, the difference was not highly significant (a = .14).

2. While students taking a-u-c quizzes during the term reported less anxiety toward test taking than students taking standard multiple-choice, the difference was not significant (a = .26).

3. There was no significant difference in the reliabilities of the a-u-c and standard multiple-choice test. Both types of quiz, however, proved to be reliable testing instruments.

Results not Related to Hypotheses

4. Students electing to take independent study Mth 101 scored significantly (a = .05) higher on the initial placement test than students electing to take the course by a standard classroom procedure.

5. There was a high variability between the three scores assigned to each student's test on the open-ended tests used in the study.

V. SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

This chapter will be divided into three sections. The first section will be a summary of the study including a review of the purpose, the design, and the results. The second section will discuss the conclusions which may be drawn from the results of this study. The third section will make recommendations for further studies.

Summary

This study was an investigation of the use of an answer-until-correct procedure for taking multiple-choice quizzes in the setting of an independent study course. The study compared the answer-until-correct method to the standard multiple-choice procedure with respect to student achievement and anxiety. It also compared the reliabilities of the two procedures.

The teaching value of the a-u-c method was examined using a group of 36 students in the Mth 101S Intermediate Algebra II class, Spring term of 1975. During the first half of the term 15 students took four unit quizzes using the a-u-c procedure and 21 students took the same quizzes using a standard multiple-choice procedure. At midterm an open-ended examination was given both groups. The mean scores on this examination were tested for significance using a standard t-test. The a-u-c group's mean was 13.5 (out of 200) higher than the multiple-choice group's which was significant at a .14 level.

The same two experimental groups were used to examine the anxiety felt by students toward the two testing procedures. Before students took their midterm examination they were asked to rate their anxiety toward the course quizzes on a five-point Likert-type scale. The means of the two groups were then tested for significant difference using a standard t-test. While the a-u-c group showed less anxiety on the average, the difference was only significant at a .26 level.

Forty-four students enrolled in Mth 95S, Intermediate Algebra I, at O.S.U. Spring term of 1975, were used to examine the comparable reliabilities of the two testing procedures. A midterm examination was written which consisted of 15 open-ended objective questions and 15 multiple-choice questions. The two sets of questions paralleled each other with respect to content and difficulty. The students were randomly assigned to four groups at the time of their midterm examination. Group 1 took the multiple-choice section using the a-u-c method and then took the open-ended section. Group 2 took the open-ended section first and then the multiple-choice section using an a-u-c procedure. Group 3 took the multiple-choice section by a standard procedure and then took the open-ended section. Group 4 took the open-ended section first and then the multiple-choice section by a standard procedure.

In examining the results, a high mean on Group 1's open-ended section suggested an interaction caused by the teaching effect of the a-u-c examination which had been taken first. This interaction was found not to be significant so was ignored. Groups 1 and 2 were combined to form an a-u-c group and Groups 3 and 4 were combined to form a standard multiple-choice group. The correlations between the scores on the multiple-choice section and the open-ended section of the test was calculated for each group. These correlations were then transformed into z-scores and tested for significant difference using the Normal Distribution. The difference between the two reliabilities favored the a-u-c group but was very small and not significant.

Another test of the comparable reliabilities of the two groups was made by rescoring the multiple-choice section of the a-u-c group above using the standard procedure. The correlation between these scores and the open-ended section scores was then calculated and compared to the correlation using the a-u-c scores. This time the difference between the two reliabilities was more pronounced in favor of the a-u-c method but still only significant at a .21 level.

Two other interesting results came out of the study. One was that students electing to take the independent study course did significantly better (a = .05) on the initial placement test for the course than those electing to take the course by a standard classroom method.

The second result was that there occurred a high variability between the scores given each student on the open-ended test questions by the three instructors grading them. The highest deviation between high and low scores was one of 53 points (on the 200 point test).

Conclusions

This study investigated the use of the a-u-c format on multiple-choice quizzes in independent study pre-calculus mathematics course. It compared quizzes in this format to those in a standard multiple-choice format with respect to teaching effect, anxiety production, and reliability. The a-u-c format proved to be at least as good if not slightly better in each area.

Therefore, this investigator concludes that this study gives just cause for the continued and expanded use of a-u-c quizzes in independent study pre-calculus courses. They have proved themselves at least equal in the major areas of concern to the most commonly used testing format in such courses. In addition, they possess the self-scoring feature which makes them more efficient to use in a learning center situation. Also, this format is very well-liked by students.

Recommendations

There are several suggestions this investigator would make to any who would reiterate this study in part or totality. They are:

1. Further study needs to be made concerning the optimal method of scoring the a-u-c format. On this study, the point count given getting the correct response on the first through fourth response was ten, five, two and zero, respectively. Upward or downward movement of this scale may improve the reliability of student scores.
2. In the reliability section of this study, students were using the a-u-c procedure for the first time. Familiarity with this format beforehand might allow students to use the format more effectively.
3. Achievement of students using a-u-c quizzes in this study was investigated after only half a term. A study which investigates student knowledge of material at the end of an entire course or their retention of this knowledge several years later would give useful additional information about the a-u-c format's teaching effect.
4. Anxiety is a complex concept and this study did nothing to differentiate between its different forms. A more thorough investigation might ascertain more information about the different types of anxiety present in students toward testing.

There are at least two further studies which are suggested by this study. One study would be a comparison of the reliabilities of open-ended objective tests and a-u-c tests. Most mathematics instructors have long held fast to the belief that open-ended tests are the most valid and reliable method of testing and yet this study indicates that such tests are subject to great variability in scoring.

The a-u-c quizzes were used in this study in a program where students could seek help after their completion of a quiz but this was not mandatory. The helpers were not other course members as in the Keller plan but rather upper class mathematics majors, graduate students or professors. These helpers were on hand in the learning center to aid students from a variety of courses. A second study would be to investigate if achievement is significantly improved if the a-u-c quizzes are used in conjunction with a proctor system where students have to review any missed problems with a helper assigned specifically to that course. Since the a-u-c quizzes provide some of the immediate feedback, which the proctorial system was designed in part to give, it should be determined how important to learning the one-to-one contact is.

BIBLIOGRAPHY

Angell, G.W. and M E. Troyer. A new self-scoring test device for improving instruction. School and Society 67:84-85. January 31, 1948.

Angell, G.W. The effect of immediate knowledge of quiz results on final examination scores in freshman chemistry. Journal of Educational Research 42:391-94. January 1949.

Gage, N.L. (ed.) Handbook of research on teaching. Chicago, Rand McNally, 1963. 1218 p.

Gilman, David and Paula Ferry. Increasing test reliability through self-scoring procedures. Journal of Educational Measurement 9:205-207. Fall 1972.

Glaser, Robert (ed.) Teaching machines and programmed learning Il, data and directions. Washington, National Education Association of the United States, 1965. 831 p.

________________. Christmas Past, present and future. Contemporary Psychology 5:24-28. 1960.

Gronland, Norman. Measurement and evaluation in teaching. New York, Macmillan Company, 1971. 545 p.

Hanna, Gerald. Improving reliability and validity of multiple-choice tests with an answer-until-correct procedure. Paper presented at joint session of American Educational Research Association and the National Council of Measurement in Education, Chicago, Illinois. April 15-19, 1974.

Harding, Larry, Phyllis Salop and Kirk Johnson. An evaluation of the use of chemically treated answer sheets. Navy Personnel Research and Development Center NPRDC TR 74-9. March 1974.

Keller, F.S. Untitled article. The Personalized System of Instruction Newsletter 1:1. June, 1971.

Keller, Fred S. and J. Gilmour Sherman. The Keller plan handbook. Menlo Park, California, W.A. Benjamin, Inc., 1974. 99 p.

Little, J.K. Results of the use of machines for testing and for drill upon learning in educational psychology. Journal of Experimental Education 3:45-49. September 1934.

Lumsdaine, A.A. and Robert Glaser (ed.) Teaching machines and programmed learning, a source book. Washington, National Educational Association of the United States, 1960. 724 p.

Peterson, J.C. A new device for teaching, testing and research learning. Transactions of the Kansas Academy of Science 33:41-47. 1930.

Peterson, Roger G. Exercises in statistical inference. Corvallis, Oregon, Oregon State University Book Stores Inc., 1972. 258. p.

Pressey, S.L. A simple apparatus which gives test and scores - and teaches. School and Society 23:373-76. March 20, 1926.

________________. A third and fourth contribution toward the coming "industrial revolution" in education. School and Society 36:668-72. November 19, 1932.

________________. Development and appraisal of devices providing immediate automatic scoring of objective tests and concomitant self-instruction. Journal of Psychology 29:417-47. April 1950.

Sherman, J.G. Questionnaire results. The Personalized System of Instruction Newsletter 2:5. June 1974.

Silver, Jerry and Bert Waits. Multiple-choice examinations in mathematics, not valid for everyone. The American Mathematical Monthly 80:937-942. October 1973.

Skinner, B.F. Teaching machines. Science 128:969-77. October 24, 1958.

________________. The technology of teaching. New York, Appleton-Century-Crofts, 1968. 482 p.

Snedecor, George and William Cochran. Statistical methods. Ames, Iowa, The Iowa State University Press, 1967. 593 p.

Strang, Harold R. and James O. Rust. The effects of immediate knowledge of results on task definition on multiple-choice answering. The Journal of Experimental Education 42:77-80. Fall 1973.

Womer, Frank Burton. The evaluation of item selection techniques appropriate to a new response method for multiple-choice test type items. Ph.D. thesis. Ann Arbor, University of Michigan, 1956. 167 numb. leaves. (Abstracted in Dissertation Abstracts 16:98. 1956)

APPENDICES

APPENDIX A

Midterm Examination Used in Achievement Study

Name ______________________________

MIDTERM EXAMINATION
MTH 101S

SHOW ALL YOUR WORK ON THE PROBLEM BELOW:

1. Rationalize the denominator: $\frac{\sqrt{x} - 5}{\sqrt{x} + 3}$

2. Simplify: $(25x^4)^{-3/2}$

3. Express as a single radical: $\sqrt{a^3} \cdot \sqrt[3]{a}$

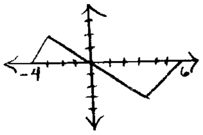
4. Convert 30 ft.^{2}/sec. into yd.^{2}/min.

5. Find the Cartesian product A X B of the set A = {1, 2} and B = {1, 5, 7}.

6. If $f(x) = 2x^3 + |x|$, find $f(-3)$.

7. If $f(x) = 3x + 2$, find $f^{-1}(8)$.

8. The graph of $y = f(x)$ is given below. Sketch the graph of $y = f(x) - 2$.



9. Find an equation of the line which passes through the points (3, 11) and (-2, 1). Express your answer in the Slope-Intercept form, $y = mx + b$.

10. Find the distance between the points (2.3) and (7, 9).

11 Sketch the graph of the parabola $y = 2x^2 + 6x + 1$. Mark the coordinates of the vertex.

12. Graph the following set on the number line.
$\{x:\ 2x+1>3\} \cap \{x:\ 5-x>1\}$.

13. Graph the set $\{x:\ |2x-3|>9\}$ on the number line.

14. Solve the following system of equations for x and y.

$$\begin{aligned} 2x - 4y &= 2 \\ 5x + 26 &= 41 \end{aligned}$$

15. A set of coins consisting of nickels and dimes is worth $2.05. If there are 23 coins in all, how many of each type of coin is in the set?

16. Solve the following system for x, y, and z.

$$\begin{aligned} x + y + z &= 7 \\ 2x - y + 3z &= -2 \\ 3x + 2y - z &= 20 \end{aligned}$$

17. The sum of three numbers is four. The third number equals -3 times the sum of the first two numbers. The second number is four less than the sum of the first and third numbers. Find the three numbers.

18. Graph the following system of inequalities on the Cartesian plane.

$$\begin{aligned} 2x - y &\geq 4 \\ 3x + 26 &\geq 6 \end{aligned}$$

19. Given that the midpoint of a segment is (0,6) and one of its endpoints is (2,3), find the other endpoint.

20. Simplify: $\sqrt{27a} + \sqrt{12a}$.

APPENDIX B

Scale Used in Anxiety Study

WE ARE INTERESTED IN DETERMINING HOW THE FORMAT OF YOUR QUIZZES IN THIS COURSE AFFECT YOUR NERVOUSNESS OR ANXIETY TOWARDS THEM. PLEASE MARK ON THE SCALE BELOW, YOUR GENERAL ANXIETY LEVEL WHILE YOU ARE TAKING MTH 101S QUIZZES.

1	2	3	4	5
Very Calm	Relatively Calm		Anxious	Very Anxious

WHAT GROUP ARE YOU IN? ____A ____B

APPENDIX C

Midterm Examination Used in Reliability Study

Name____________________________

MIDTERM EXAMINATION
MTH 95S - PART ONE

SHOW ALL YOUR WORK ON THE PROBLEMS BELOW:

1. What is the decimal representation of $\frac{14}{111}$?

2. Evaluate $xy^3 - yz^2$ when $x = 5$, $y = -2$ and $z = 4$.

3. Simplify: $2\{5-4[2x-(5x+4)]\}$.

4. Solve for x: $3(x+2) = 4 - (x-6)$.

5. A family pays $400 for a washer and dryer. The washer costs $30 more than the dryer. Write an equation or system of equations whose solution will yield the cost of each unit. Use x for the cost of the dryer. You do not need to solve the equation or equations.

6. Solve for h: $a = 2(h-b)$.

7. What is the y-intercept of the line $5x - 2y = -15$?

8. Find the equation of the line passing through the points $(6,4)$ and $(-3,2)$.

9. Solve the following system of equations for x and y.

$$\begin{aligned} 3x + 2y &= 2 \\ 2x - 3y &= -16 \end{aligned}$$

10. Solve the following system of equations for x, y, and z.

$$\begin{aligned} x + y + z &= 9 \\ 2x - y + 3z &= 3 \\ 3x - 3y - z &= -3 \end{aligned}$$

11. The sum of Bill and Barry's ages is 23. In five years Bill will be twice as old as Barry will be then. How old is each now?

12. Multiply $x^3 - 3x + 5$ by $x-3$.

13. Factor completely: $6x^2 - 5x - 4$.

14. Solve for x: $x^2 = 4x + 12$.

15. On the number line, graph the solution set of $|x+1| < 3$. Be sure to indicate your coordinates.

Name____________________________

MIDTERM EXAMINATION
MTH 95S - PART TWO

MULTIPLE CHOICE

1. Which of the following is the decimal representation of $\frac{5}{11}$?

   ___A. $.\overline{45}$ ___B. $.4\overline{5}$ ___C. $.45$ ___D. None of these

2. What is the value of $-a^3b + cb^2$ when $a = -3$, $b = 2$, and $c = -1$?

   ___A. -58 ___B. -18 ___C. 14 ___D. 50

3. Simplify: $5 + 3[2-x(3-x)+2x]$

   ___A. $16-8x+8x^2$ ___B. $11-3x+3x^2$ ___C. 11 ___D. $16-8x-8x^2$

4. Solve for a: $3 - 2(a+6) = 5 - a$

   ___A. $a = -14$ ___B. $a = 4$ ___C. $a = -14/3$ ___D. $a = -4$

5. Three consecutive even integers have a sum of 72. If x is used to represent the smallest integer, which equation below could be used to find the three integers?

   ___A. $x+(x+1)+(x+2) = 72$ ___B. $x+2x+3x = 72$

   ___C. $x+(x+2)+(x+4) = 72$ ___D. $x+2x+4x = 72$

6. Solve for s: $2s - a = cs$

   ___A. $s = \frac{c+a}{2}$ ___B. $s = \frac{a}{2-c}$ ___C. $s = \frac{c+a}{3}$ ___D. $s = 2+c+a$

7. What is the x-intercept of the line $4x+3y-12 = 0$?

   ___A. (3,0) ___B. (0,4) ___C. (-3,0) ___D. (0,-4)

8. What is the equation of the line which passes through the points (2,5) and (-3,10)?

   ___A. $y = -x+3$ ___B. $y = x+3$ ___C. $y = -x+7$ ___D. None of these

9. Solve the following system of equations for x and y.

$$4x - 2y = -13$$
$$x + 3y = 9$$

The <u>sum</u> of x and y equals

___A. 2 ___B. 5 ___C. 10 ___D. None of these

10. Solve the following system of equations for a, b, and c.

$$a + b + 2c = 9$$
$$2a + 3b + c = 5$$
$$a - 2b - c = 0$$

Which of the following is true?

___A. $a = 2$ ___B. $b = 1$ ___C. $c = 5$ ___D. $a = 4$

11. One adult ticket and one student ticket to a given movie together cost $4. 25. Three adult tickets and two student tickets altogether cost $11.00. What is the price of an adult ticket?

___A. $1.50 ___B. $1.75 ___C. $2.00 ___D $2.50

12. When x^2-x+5 is multiplied by x^2-2x, what is the coefficient of the x^2 term?

___A. 1 ___B. 2 ___C. 3 ___D. 7

13. One of the factors of $10x^2-13x-3$ is

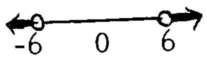
___A. $10x+1$ ___B. $2x-1$ ___C. $5x+1$ ___D. $2x+3$

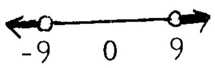
14. Solve for x: $x^2-10 = 3x$

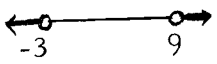
___A. $x = 5$ or $x = -2$ ___B. $x = 5$ or $x = -5$

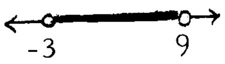
___C. $x = -5$ or $x = 2$ ___D. None of these

15. Which of the following is the graph of the solution set of

___A. 

___B. 

___C. 

___D. 

APPENDIX D

Item Analysis of Mth 95S Midterm Examination - Part Two

QUESTION NUMBER	DIFFICULTY INDEX	DISCRIMINATION INDEX
1	96	8
2	67	67
3	79	42
4	75	50
5	88	25
6	71	58
7	71	58
8	67	67
9	48	45
10	62	57
11	96	8
12	75	50
13	83	33
14	92	17
15	63	75

APPENDIX E

Scores on Mth 95S Midterm Examination Open-Ended Section

Student (Coded)	Scorer 1	Scorer 2	Scorer 3	Mean	High-Low Score
1	144	140	140	141	4
2	104	115	100	106	15
3	119	127	115	120	12
4	79	106	75	87	31
5	100	95	101	122	14
6	56	67	60	61	11
7	150	141	150	147	9
8	122	125	125	124	3
9	77	95	75	82	20
10	136	134	135	135	2
11	134	138	130	134	8
12	122	133	110	122	23
13	140	146	140	142	6
14	144	145	140	143	5
15	47	60	35	47	25
16	95	100	95	97	5
17	140	138	140	139	2
18	65	76	60	67	16
19	80	103	70	107	33
20	122	127	110	120	12
21	150	150	140	147	10
22	109	122	100	110	22
23	127	127	125	126	2
24	104	121	100	108	21
25	97	120	100	106	23
26	100	119	110	110	19

Student (Coded)	Scorer 1	Scorer 2	Scorer 3	Mean	High-Low Score
27	60	82	55	67	27
28	49	72	20	47	52
29	91	94	80	88	14
30	85	91	85	87	6
31	105	133	85	108	48
32	96	109	85	97	24
33	59	74	65	66	15
34	135	146	135	139	11
35	79	94	70	81	24
36	112	125	105	114	20
37	45	72	55	57	27
38	140	138	135	138	5
39	72	80	60	71	20
40	108	106	95	101	13
41	77	81	75	78	6
42	74	72	65	70	9
43	80	83	85	83	5
44	150	148	150	149	2

APPENDIX F

Scores on Mth 101S Open-Ended Midterm Examination

Student (Coded)	Scorer 1	Scorer 2	Scorer 3	Mean	High-Low Score
1	185	196	175	185	21
2	141	167	160	156	26
3	184	185	180	183	5
4	102	128	75	102	53
5	162	185	160	169	25
6	109	144	105	119	39
7	39	50	40	43	11
8	132	148	135	138	16
9	77	97	85	86	20
10	149	163	150	154	14
11	124	158	120	134	38
12	177	192	185	185	15
13	130	160	140	143	30
14	130	150	140	140	20
15	122	144	120	129	24
16	177	192	185	185	15
17	100	122	110	111	22
18	71	92	70	78	22
19	108	119	100	109	19
20	83	88	80	84	8
21	127	141	130	133	14
22	94	122	95	104	28
23	149	164	150	154	15
24	86	117	85	96	32
25	113	74	115	101	41
26	91	109	95	98	18

Student (Coded)	Scorer 1	Scorer 2	Scorer 3	Mean	High-Low Score
27	148	170	150	156	22
28	192	199	195	195	7
29	27	27	25	26	2
30	88	106	80	91	26
31	147	161	135	148	26
32	99	121	85	102	36
33	158	175	155	163	20
34	159	190	160	170	31
35	163	167	165	165	4
36	112	139	90	114	49

APPENDIX G

Sample Mth 101S Quiz

Mth 101S Name______________________________

UNIT III
QUIZ

1. Completing the square, the function $y = x^2+5x-3$ becomes

   _0_ A. $y = (x+5/2)^2 - 11/2$ _+_ B. $y = (x+5/2)^2 - 37/2$

   _0_ C. $y = (x+5)^2 - 8$ _0_ D. None of these

2. The vertex of the parabola $y = -x^2+2x+4$ is

   _+_ A. (1,4) _0_ B. (-1,3) _0_ C. (1,3) _0_ D. (-1,5)

3. Which of the following is true concerning the function $f(x) = x^2+4x+3$?

   _0_ A. The maximum value of $f(x)$ is 3.

   _0_ B. The minimum value of $f(x)$ is 7.

   _0_ C. The maximum value of $f(x)$ is -1.

   _+_ D. The minimum value of $f(x)$ is -1.

4. What are the x-intercepts of the function $y = x^2+5x-24$?

   _0_ A. (8,0), (-3,0) _0_ B. (6,0), (-4,0)

   _+_ C. (-8,0), (3,0) _0_ D. (5/2,0), (121/4,0)

5. Which of the following is the graph of $y = \sqrt{x-3}$?

   _0_ A. _0_ B. _+_ C. _0_ D.

6. If a = {1, 5, 6, 9, 10} and B = {3, 4, 5, 6, 7} then A $\cap$ B =

0 A. Ø + B. {5,6} 0 C. {1, 3, 4, 5, 6, 7, 8, 9, 10}

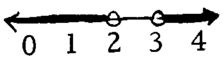
0 D. {1, 3, 4, 7, 9, 10}

7. Solve for x: $\frac{x}{2} + 3 < 2x - 3$

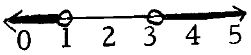
0 A. $x < 6$ 0 B. $x > 0$ 0 C. $x > 6$ + D.. $x > 4$

8. Which of the following is the graph of $\{x: 2x-1<3\} \cup \{x: 3x>9\}$

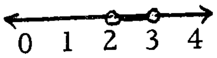
+ A.



0 B.



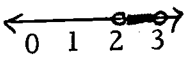
0 C.



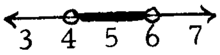
0 D. None of these

9. Which of the following is the graph of x: 2 2x-2 4 ?

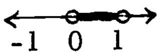
+ A.



0 B.



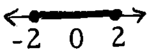
0 C.



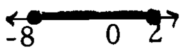
0 D. None of these

10. Which of the following is the graph of $|x+3| \leq 5$?

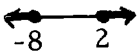
0 A.



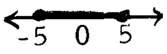
+ B.



0 C.



0 D.



The pluses and zeros which appear before each response are latent images on the quiz which are brought to view when a student marks the area with the special marker.