THE EFFECTS OF VARIOUS CONDITIONS OF ADJUNCTIVE INTERACTIVE COMPUTER-ASSISTED TESTING ON FINAL EXAMINATION PERFORMANCE

by

Joshua Hausman

A THESIS PRESENTED TO THE
FACULTY OF GRADUATE STUDIES AND RESEARCH, MCGILL UNIVERSITY,
IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF ARTS
DEPARTMENT OF EDUCATIONAL PSYCHOLOGY AND SOCIOLOGY
JULY, 1978

SHORT TITLE

INTERACTIVE COMPUTER-ASSISTED TESTING

JOSHUA HAUSMAN

MASTER OF ARTS

DEPARTMENT OF EDUCATIONAL PSYCHOLOGY AND SOCIOLOGY

ABSTRACT

The study explores the effects of various conditions of interactive computer-assisted testing (ICAT) on final examination performance. One hundred and five subjects were randomly assigned to one of three treatments. Control group 1 was scored on a right-wrong basis with no item retrys. Experimental group 2 was scored with fractional marks given for item retrys. Experimental group 3 was scored on a right-wrong basis without credit given for required item retrys.

Results suggest that partial scoring is more predictive of final examination performance scores. The study supports the idea that higher mastery levels on ICAT quizzes are associated with superior criterion performance.

RESUME

Cette recherche étudie les effets dans des conditions variables d'examens donnés par ordinateur (ICAT) sur le rendement d'un examen final. Cent cinq sujets ont été sélectionnés au hasard et soumis à un des trois traitements. Le groupe contrôle 1 fut noté sur une base de vrai ou faux sans aucune reprise possible. Le groupe expérimental 2 fut noté avec des décimales pour signifier les essais supplémentaires. Le groupe expérimental 3 fut noté sur une base de vrai ou faux sans aucune attribution pour les essais supplémentaires de certaines questions.

Les résultats suggèrent que la correction partielle prédit avec plus de précision le rendement et les notes de l'examen final. Cette étude soutient l'idée que le niveau de maîtrise des épreuves du "ICAT" est directement lié au critère de rendement supérieur final.

ACKNOWLEDGEMENTS

The author is grateful to many who have helped during the course of writing the thesis. Thanks to Mrs. Lois Norton, Dr. Socrates Rapagna, Dr. Jeffrey Derevensky for their aid in completing the thesis. Special thanks to Charles Quesnel and Pierre Parent for help with computer programing.

This research was supported in part by a grant from F.C.A.C., Ministère de l'Education, Gouvernement du Québec, to whom I express my gratitude.

Finally, to my advisor, Dr. Glenn F. Cartwright, goes indebted thanks and appreciation for his skillful support and vigilant guidance which gave me encouragement and resulted in an interactive learning experience which will be hard to forget.

J.H.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER I

THEORETICAL FRAMEWORK OF THE STUDY

Interactive Computer Usage

As computer technology becomes a less expensive and more useful technology in educational environments, teachers will rely less on traditional classroom lectures for teaching and paper-and-pencil tests for evaluation. With computer-assisted instruction (CAI) exists the potential for truly individualized instruction. No apparent educational sacrifice is inherent with CAI, either in overall control of the management of the instruction or the flexibility needed for creative teaching. CAI gives the teacher a more diversified and powerful medium for teaching materials (Cohen, 1975).

Interactive Computer Testing

The computer can be used in the construction of examinations. Computer-assisted test construction (CATC) is a system designed to select questions with certain attributes, organize them into a test and print it in an error-free format (Cartwright, 1975). The multiple-choice format is chosen as a more objective form of testing and can be done by machine. Each student responds to the multiple-choice test by recording his responses on a machine readable answer card. This answer card is later scored by the computer with student's scores and statistical data available later on computer printouts. While CATC appears to be an effective, low cost procedure for use in evaluation of student performance, it lacks immediate feedback. Often there is a long delay between the writing of the test and the return of the results.

The addition of immediate feedback is possible with ICAT which allows the computer to administer the test and give immediate feedback at the end of the test or even after each question. In addition the feedback can range from immediate knowledge of results (KR), to knowledge of correct results (KCR), to elaborate explanatory information paragraphs, or any combination of the above.

The provision of immediate feedback ensures to some degree that students learn while taking each quiz. Because ICAT uses on-line computational facilities, further development of more sophisticated forms of tailored testing basing item selection on student characteristics and performance and/or item attributes, is possible (Cartwright & Derevensky, 1977).

Franklin and Marasco (1977) examined interactive computer-based testing (ICBT) in university level science courses to study its effectiveness as an educational tool. ICBT is similiar to ICAT but it is a term that Franklin and Marasco (1977) use when referring to interactive computer testing. They discuss some of the myths regarding its use in education. They contend it is untrue that requiring a student to use computer facilities is an obstacle in pursuit of his educational goals. Rather it allows for a variety and flexibility heretofore available only in oral examinations. This form of testing results in considerable amount of student-teacher interaction and an increase in informational value of the test with individualized, immediate, detailed and constructive feedback.

Franklin and Marasco (1977) point out that the construction of examinations is a task which many instructors find difficult and which demands a lot of time. Publishers have attempted to ease the burden of making examinations by providing instructors guides containing test items. The instructor looks at the guide and has a typist prepare the examination. Short-answer or essay questions may require many hours of time to evaluate and often the task is given to teaching assistants or 'readers'. Instructors without such help often choose multiple-choice examinations because of scoring ease.

Cost and Other Problems

In terms of cost, any labour intensive approach to testing will ultimately be more expensive than an approach based on technologies which each year deliver more per dollar. Franklin and Marasco contend that this form of testing does not have significant problems with security and cheating, since the problem may be resolved by controlling access to files and programs. Now that test programs randomly select questions from a large item pools, tests can be reassembled without worrying that students have seen the test.

Sophisticated Tactics

ICBT provides a variety of sophisticated tactics. The computer is ideally suited for the solution of complex problems, by breaking each down to its component parts. The computer may decide not to reveal the next part of a question if the first part was not successfully completed: a control option not present in paper-and- pencil tests.

The computer can also perform different routines for multiple-choice questions. For example, the computer may withhold the display of certain alternatives until the student rejected the earlier ones. This type of process is less vulnerable to elimination strategies by the student.

The computer may present a problem with insufficient information, telling the student that additional information will be supplied upon request. The student learns that an important step in solving a problem is to ask oneself what data are required.

The computer can be used to simulate physical systems. The student interacts with the system by answering questions, observing the response and then deciding the underlying principle or concept.

ICBT is amenable to virtually any pedagogical strategy. Instructors can vary difficulty of exams for different audiences, type of question, exam length, and exam ending. Besides providing feedback, it has extensive record keeping and clerical capabilities. By singling out those students who have difficulty in particular areas, the record keeping portion of ICBT allows an instructor to help those students who need help.

For a few years, the University of California at Irvine, has run its large pre-calculus course with all unit by unit tests using interactive computer-administered quizzes. A similiar course with approximately the same number of students (several hundred) is also taught at the University of California, San Diego, but without using computers. Comparison shows that the funds spent on using the computer resulted in a monetary saving of twice the amount than when the computer was not used (Franklin and Marasco, 1977).

Franklin and Marasco (1977) stress that students should have, and know they have, free access to human beings to ask questions and make comments about the materials. ICBT meets existing needs and its flexibility is a definite asset in meeting the requirements of its users. The tendency towards continuing education, and the success of, and demand for individual instruction are all factors which favour ICBT.

General Statement of Problem

The present research is to investigate in what ways ICAT may improve learning. The effect of transfer of on-line performance to criterion performance will be examined. In addition a pre-test and a post-test will be included in the design to examine gain scores.

Because Evans and Misfeldt (1974) have indicated that partial scoring increased total test reliability, partial scoring for the quizzes will be examined to see if this better predicts criterion performance. Research will also look at what effect the number of explanatory paragraphs requested by students has on quiz taking behavior and criterion performance. Further, the reinforcing effects of partial scoring will be investigated to ascertian its effect on learning.

To summarize, ICAT appears to have been moving towards the refinement of both its evaluation and learning roles. New scoring procedures have modified the feedback and reinforcing role of the computer. The addition of the pre-test and the post-test with both theoretical and applied questions, will enable a closer examination of a number of variables which affect learning.

CHAPTER II

REVIEW OF RELATED LITERATURE

Overview of the Chapter

In this chapter the literature related to the present study is reviewed. This includes the research on ICAT as well as off-line computer-assisted testing and feedback. The relevant literature pertaining to computer-adaptive testing and partial scoring is summarized. Finally, a review of research relating feedback to learning and reinforcement is given.

Evolution of ICAT

Historically, the antecedent of ICAT is Pressey's (1926) teaching machine. In the 1920's, Pressey designed several machines that automatically tested a student with multiple-choice questions and provided feedback on each decision or knowledge of results (KR). If the student selected correctly he moved to the next question. If he were incorrect the error was tallied and he continued to respond until he chose the correct answer.

However, the historical trend of the machine teaching shifted in the 1950's. Pressey's teaching machine failed to gain popularity and the shift to linear programmed instruction moved the research more towards software technology where educational programs were stressed and individualized instruction emphasized. Skinner (1961) felt that each student should compose his own response based on recall, rather than select an answer from a set of alternatives, which is basically a recognition task. Further, since Skinner was interested in shaping behavior, choosing incorrect multiple-choice alternatives only served to strengthen unwanted responses. Holland and Skinner (1961) used twenty machines and programs to teach a part of a course in human behavior to undergraduates at Radcliffe and Harvard. They carefully sequenced the learning frames so that the frames would be thoroughly understood before the student moved on.

ICAT re-emphasizes hardware technology somewhat, by using the machine to exert greater stimulus control. With improved technology, computers can go beyond Pressey's teaching device. New technology such as time-sharing developed in the 1960's, allowed for an 'individualized' computer-user interaction and provision of knowledge of correct results (KCR). With this development the computer test can be more of a learning catalyst. This new technology changes our understanding of the 'test'. The methods by which the test is presented has changed with ICAT (Cartwright, 1971).

This form of computer-user interaction is an improvement on another form of computer-assisted testing (CATC). CATC aids the instructor to assemble tests (Cartwright, 1975). ICAT has two capabilities, it can provide KCR as well as help in the evaluation process. ICAT is a process in which the computer administers a test to each user and gives immediate feedback regarding the user's responses. More information may be provided to the user in the form of feedback paragraphs.

By providing KCR, ICAT is also a learning tool. The advent of ICAT is in line with some other historical trends. Before ICAT, computer assisted instruction (CAI) was seen in the role of a teacher. CAI administers student instruction by means of a computer. It accepts and processes student interaction, controls further progression, and may provide remediation. With ICAT, interactive computer testing can also assume part of the role of a teacher with its interactive provision of feedback.

At McGill University, ICAT was implemented in 1970 in an introductory psychology course. The program consisted of six multiple-choice exams. The test items were coded in a CAI author language known as MULE (McGill University Language of Education) by sixty students each of whom programmed approximately fourteen questions to produce an item bank of over 800 questions. The project was deemed to be successful (Cartwright, 1971) and students seemed to have benefitted from the addition of ICAT to the course. Consequently, research continued to develop and improve ICAT. In 1973, work was completed coding items from an introductory educational psychology text for a computer presentation. The questions were provided by the publisher. Immediate KCR was provided and elaborate feedback paragraphs were incorporated. These explanatory paragraphs indicated to the student the correct answer, why alternatives were incorrect, and contained page numbers for further reference. The items were coded in a CAI author language known as CAN-VI (Cartwright & Tessler, 1975) and were processed by an IBM 370 series computer and on ten Teletype terminals which were later replaced by cathode-ray tube (CRT) terminals.

Five quizzes were established based on specific chapters of the textbook (Biehler, 1971, 1974). In 1976, a sixth quiz was added at the students' request, to sample questions from the entire text. Each quiz consists of twenty questions based on specific textbook chapters which are randomly selected from the bank of items. After each question item, KCR is provided and the opportunity is given for students to request a feedback paragraph. In addition a utility function was developed to allow students to send and receive messages from their instructor, to use the computer for calculation, and to see their current score file (Derevensky & Cartwright, 1977).

The quizzes are used as adjuncts to traditional classes. Students towards the end of the year began perceiving the quizzes more as a learning than an evaluative tool. It is possible that the feedback paragraphs contributed to the students changed perception of the quizzes (Cartwright & Derevensky, 1976). Research was undertaken to compare the effects of exposure to computer-assisted testing (CAT) to students who were not exposed to CAT (Cartwright and Derevensky 1975). One hundred and twenty-four subjects were divided into six sections of an introductory course in educational psychology with three sections exposed to more traditional type of evaluation. The Teaching Methods Questionaire II was administered (Cartwright, 1973), to determine if differences in attitudes existed between CAT and non-CAT groups. Students exposed to CAT tended to perceive the computer quizzes as being more of a learning than an evaluative experience, reported learning more from computer quizzes than traditional classroom examinations, and in general tended to rate CAT as being superior to traditional classroom examinations.

In one study, it was found that students who had achieved a 90-100% average score over the five quizzes scored approximately 10% higher on the final criterion test than students who had achieved only 80-90% on the quizzes which suggested that the higher the criterion level for mastery, the greater the positive transfer effect. The total amount of time spent was not significantly related to final criterion performance. The number of times quizzes were taken was only weakly correlated with final criterion performance (Cartwright & Derevensky, 1977).

To summarize, ICAT allows the student to sit at a computer terminal and be administered randomly chosen multiple-choice questions. The chief advantage of ICAT (Cartwright & Derevensky 1976) is its on-line nature permitting immediate feedback during the test. A current disadvantage with the on-line procedure is its high cost

(Lippey 1977). However with technological advances the cost will be reduced and minicomputers might replace the use of expensive computers. ICAT has other potentials due to its on-line nature permitting other sophisticated testing procedures to be added. The following summarizes some of the types of testing procedures that are now available and which can be eventually incorporated within ICAT.

Computer Testing

Strategies of Adaptive Ability Measurement

For almost sixty years, the predominant mode of administration of ability tests has been paper-and-pencil multiple-choice tests. These multiple-choice tests are often too difficult for some people and too easy for others. If the test is too difficult, the testee may become frustrated; if the test is too easy, the testee may become bored and may not put in maximum effort. These possibilities introduce error which may lower the reliability of test scores. Theoretical studies comparing conventional and tailored (adaptive) tests show that the closer the probability of correct responses for each item is to 50% the greater the accuracy of the test score (Lord 1970, 1971a,c,d,e). Adaptive testing, is where the test items are adapted to the ability level of the testee and not to the average ability level of the group. The process of adapting testing is done using responses to earlier test items which enables a selection of items that have a 50% level of difficulty for the testee. Consequently, improvement is made in testing strategies by having the number of items above or below the subject's ability minimized to reduce error variance.

Other problems with paper-and-pencil multiple-choice tests is the use of time limits which may introduce error in test measurement because individuals respond differently to time pressures. Testing without time limits allows testing to be more catered to the individual needs (Weiss & Betz 1973). Administration-testee interactions may increase error variance even under group testing conditions (Weiss & Betz, 1973).

In previous research on adaptive testing, tests were administered using paper and pencil, testing machines or computers. Now computerized adaptive testing ability tests may be redesigned for administration to each testee on CRT terminals, teletypewriters or slide projector screens, connected to an on-line computer system (Dewitt & Weiss, 1974). The response to each test item is immediately scored by the computer, and the next question is chosen from an item pool by the computer program according to a specified adaptive testing strategy and presented for the testee's response. Different strategies represent different ways of moving a testee through an item pool.

The two stage test is the simplest of the adaptive testing strategies. It usually consists of a routing test, and a measurement test. The routing test is a short test to make an initial estimate of the individual's ability level from which he is branched to a measurement test based on his score. A number of variations have been proposed for two-stage testing strategies (Cleary, Linn, & Rock, 1968; Linn, Rock, & Cleary, 1969).

The obvious advantage to the two stage test is its adaptiveness. Since the routing test is usually relatively short in comparison to the measurement test it will provide more information per item, over more items, and thereby reduce the negative psychological effects of a conventional test which may be either too difficult or too easy for a given testee. There are two limitations to this form of testing; routing errors can be made in the assignment of measurement tests due to individuals whose scores fall near the cut off points established for assignment to different measurement tests. Also, the administration of the test requires that testees answer all items on both the routing and measurement tests.

The majority of research in adaptive testing has been fixed branching multi-stage testing strategies (Weiss & Betz 1973). These strategies differ from the two-stage test in that more than one branching decision (routing to measurement test) is made. The fixed-branching multi-stage strategies use the same item pool structure for all individuals but individuals move through the structure in different ways. A branching rule is specified prior to testing and this rule determines how an individual moves from an item at one stage of testing to an item at the next stage of testing. The branching rule in conjunction with information on whether the testee answered a given item correctly or incorrectly determines how the testee moves through the structured item pool.

The pyramidal model is a model which uses the multi-stage strategy. The pyramidal models include many variations and these can be differentiated into those using constant step sizes, variably decreasing step sizes, truncated pyramids, multiple-item pyramids, and pyramids using differential response option branching. Pyramidal tests use all items in the rating procedure and the measurement simultaneously. For example in constant step size pyramids, movement through the pyramidal structure begins for all testees at the first stage. The response to the item at stage 1 is scored as correct or incorrect, the branching rule is consulted and the appropiate stage 2 item is then administered. A typical branch rule is up-one, down-one. Using this branching rule following a correct response to an item the testee receives an item one increment higher in difficulty. Following an incorrect response he is branched to an item one increment lower in difficulty (a slightly easier item). Pyramidal tests have the capability of estimating a testee's ability in as few as 10 or 15 items, approximating the number that a two stage strategy requires for a routing test. In general the pyramidal strategies requires less items in the item pool than two-stage strategies. However pyramidal models have a recoverability problem: chance successes or failures due to irrelevant ability which may prevent the testee from going on to a higher path or a lower path. One objective of adaptive testing is to permit the tester to administer items that converge upon a difficulty level that is appropiate for each testee. In pyramidal testing this is only done for testees of about average ability.

Different forms of adaptive tests will prove that on-line testing is a more powerful means of testing. A possible addition to an on-line system is the flexilevel test. Lord (1971b) proposed the flexilevel test which is a modified pyramidal adaptive test. The flexilevel test consists of one item at each of a number of equally spaced difficulty levels. The flexilevel item structure is different from the typical pyramidal models in that the pyramidal modes have more than one item at each difficulty level. In the flexilevel test, the branching rule states that following a correct response the next item given is the item next higher in difficulty which had not previously been administered and following an incorrect response the testee receives the item next lower in difficulty that has not previously been administered.

The flexilevel test has a number of advantages over other computer adaptive testing models. Like the two-stage test, it might be possible to administer a flexilevel test by paper and pencil. It is easy to score and requires a smaller item pool, a 10 stage pyramidal test requires a 55 item structure while a 10 stage flexilevel test requires only 19 items.

There are a number of limitations with the flexilevel test. The item difficulty level diverges from the testees ability level. Such divergence might have a detrimental effect on testee motivation and result in guessing behavior on items that are too difficult. Only one item at each difficulty level can not accurately determine a testee's ability status with a high degree of precision. Consequently, flexilevel tests might be more unstable than other testing strategies if guessing is possible.

Another capability of an on-line system is the stradaptive (stratified-adaptive) computerized test. This test (Weiss, 1973) operates from an item pool in which test items are grouped into ten levels or strata according to their difficulty. Each stratum can be thought of as a peaked test in which the items are clustered around some average difficulty level. The strata are arranged in increasing order of difficulty. Unlike the other fixed branching models, stratadaptive testing begins with an estimation of the person's ability level from either prior information available on the testee or from his own self-report (Weiss, 1976). This information determines the testee's entry point. Based on whatever prior information is available, the testee of lower estimated ability begins the test with less difficult items and the testee of higher ability begins with more difficult items.

Branching in the stradaptive test occurs between stratas and any of the rules (up-one, down-one) can be used. In the up-one, down-one rule, a correct answer to an item at one stratum leads to the next available item at the stratum next higher in difficulty. An incorrect answer to an item at a given stratum branches the testee to the next available item at the stratum next lower in difficulty. The item to be administered at each stratum has not previously been administered. The stradaptive branching procedure is designed to converge upon the region of the item pool of appropiate difficulty for a given testee. In the process of this convergence, the test will locate a stratum of the item pool at which the testee answers all (or almost all) of the items correctly. This can be referred to as the basal stratum. At the same time the procedure will locate a ceiling stratum, the stratum at which the testee answers all (or almost all) the items incorrectly. In between these two strata the testee will answer 50% of the items correctly.

Stradaptive testing has several disadvantages over the other two-stage and multi-stage fixed branching models. First it is explicitly designed to take account for guessing behavior. It has complete recoverability (a chance response would not pre-empt the testee from a certain level). The test locates the region of item pool for testee. The test makes use of prior information which should act to reduce the number of items. The test varies the number of items which results in more precision. Although designed for computers it can be used on a testing machine especially designed for that purpose. ICAT with its on-line testing procedures can potentially incorporate these features. Testing will be done more intelligently because the testing procedure will take into account the earlier responses of the user, thereby introducing individualized testing. Hence on-line testing rather than off-line testing will make more advantageous use of computer technology.

Another feature of computer-delivered testing that has been incorperated into ICAT on an experimental basis is partial scoring. Partial scoring is a scoring procedure where the student continues to respond to a multiple choice question until he chooses the correct response and is scored on a 3-2-1-0 basis depending on the number of choices required to pick a correct answer assuming only four choices per item. Such scoring results in increasing the length of the examination, relaxing the scoring criteria for grades on examination and increasing test reliability (Evans & Surkan, 1977a). Partial scoring has been incorperated into computer testing (Evans & Surkan, 1977b).

Immediate Knowledge of Results and Adaptive Testing

There has been little research in the potential of computer-administered testing procedures to improve the psychological environment of ability testing. It has been suggested (eg. Hansen, Johnson, Fagan, Tam & Dick, 1974; Weiss and Betz, 1973) that adaptive testing procedures should create a more favourable psychological environment for all testees than do conventional non-adaptive tests. An approach to improving the psychological environment of testing involves providing KR. Bayroff (1964), and Ferguson & Hsu (1971) have postulated that immediate knowledge of results has an incentive or motivational effect on examinee performance on ability tests. On-line testing will prove to be an improved psychological environment for testing. The administration of ability tests by computer allows fast and efficient provision of KR to examinees.

A hypothesis concerning the relationship between the ability level and the effects of KR was offered by Betz and Weiss (1976). For high-ability students receiving high proportions of positive KR, KR would be encouraging and motivating. Low-ability examinees receiving mostly negative KR (incorrect) may be discouraged. Therefore effects of KR may be related to ability level of examinees.

A study was done (Betz, 1976; Betz & Weiss, 1976) investigating the effects of immediate knowledge of results on adaptive versus conventional testing strategies on several aspects of ability test performance and examinee behavior. The design of the study involved computerized administration of a fifty-item conventional ability test and a stradaptive ability test either with or without immediate knowledge of results. Two groups of subjects were used, one group was considered the high-ability group and the other consisted of the low-ability group.

The outcome reflected the different reactions of the high and low-ability groups to the provision of KR. In the low-ability group, mean levels of reported motivation was lower under KR. In the high-ability group mean levels of reported motivation were slightly higher under KR conditions than under non-KR conditions. While this latter difference was not statistically significant it was larger and in the opposite direction than the difference between KR and non-KR conditions in the low-ability group.

Feedback

Most feedback used in computer-assisted-instruction is one of, no feedback, feedback with KR, KCR, KCR plus an explanation, or KR plus interactive teaching (Roper, 1977). Roper's study using computer facilities presented an adjunct program in statistics. He divided examinees into three groups. Group A received no feedback, Group B received KR and Group C received KR and a sentence giving the correct answer. A post-test was given and three factors were studied. Factor x indicated number of incorrect responses on programme but correct on post-test. Factor y indicated number of correct on programme but incorrect on post-test. Factor z was the difference between factors x and y.

Results indicated a mean difference on factor x showing a significant comparison between groups A and C and between groups B and C but not between groups A and B. The results indicate that KR is not effective in correcting erroneous responses which confirms Gilman's (1970) assertion that statements such as "you are correct" are not effective (KR) as KCR. Although, Lublin (1965) indicated that KCR has a detrimental effect in programmed learning texts. Anderson, Kulhavy and Andre (1971) point out that Lublin's results was due to prompting of KCR.

Skinner is a strong proponent of KR believing that KCR does not improve learning (Roper, 1977). Guthrie (1971) wrote that KCR has the effect of correcting incorrect responses and concluded with a refutation of Skinner. However Guthrie's study appears confounded because KR and KCR were mixed together (Roper, 1977). Roper (1977) separates KR and KCR into different groups and does indicate a significant effect for KCR.

In terms of distinguishing if KCR is reinforcing or corrective, Buss, Bradel, Orgel and Buss (1956) found that either KR or no comment for a correct answer and "wrong" for an incorrect answer produced more learning than saying "right" when the responses were correct and giving no comment when it was incorrect which suggests that KCR functions more as a corrective feedback than as reinforcement.

Tait, Hartley and Anderson (1973) have suggested that in many CAI tests users were not attending to feedback. Studies in programmed instruction have suggested that delayed KCR may be more advantageous than immediate KCR (Sturges, 1972; Kulhavy & Anderson, 1972). Sturges (1977) found that delayed feedback, twenty minutes or twenty-four hours later, showed greater retention one to three weeks later than if the feedback was delayed for two seconds. However no significant difference was found between the twenty minutes and the twenty-four hour delay. If KCR is reinforcing a delay should cause interference but it has been shown to improve retention (Sasserath & Yonge, 1969). Kulhavy and Anderson (1972) have suggested that the incorrect response interferes proactively with acquisition of the correct answer (KCR). The delay allows the incorrect response to be forgotten. The delay of feedback for twenty-four hours has been shown to decrease the probability of repeating an incorrect answer on the post-test (Kulhavy & Anderson, 1972; Surber & Anderson, 1975). Also, subjects experience more difficulty in remembering their errors following a delay interval . Studies have indicated that the delay of feedback results in subjects spending significantly longer time studying the feedback (Kulhavy, 1977). Consequently, delayed feedback would probably be a more advantageous use of feedback.

Feedback acts to guide the student in learning the textual information. It also informs the student about the accuracy of his response relative to the body of knowledge that is to be learned. The use of feedback introduces a control system that regulates the learning process in an attempt to transfer information from the text to the learner (Talyzina, 1973). In a controlled system the term feedback relays information about the development of the controlled process, meaning that it is an indicator of the state of the controlled process (Talyzina, 1973). Systematic feedback allows for optimal system control in guiding the student from the initial state of the controlled process to the aim of the control process (Talyzina, 1973). If the feedback were simply a reinforcing agent one would expect intermittent feedback to be more effective.

In order to understand how feedback and learning relate to each other one must look at the systems in which it operates. There are two forms of control systems: the closed-loop control which provides feedback and regulation of the learning process and open-loop control which has no feedback. Within closed-loop control there are two types: the "black box" and the "white box" principle. The former utilizes feedback only to control the output of the learning process while in the latter, feedback provides information about the states of the transient stages of the process (Talyzina, 1973). Within the white box control system one of the requirements is the provision of the processing of information received via the feedback channel which provides correcting (regulating) actions followed by the realization of corrected responses.

Although feedback is basically a cognitive component, it still needs a reinforcement incentive model. Feedback will only increase learning efficiency when it takes into account pupil's needs to be motivated to learn (Feldhussen & Birt, 1962). To summarize, feedback can be reinforcing when the subject matter arouses interest and the acquisition of information is a satisfying experience. Consequently a control process that does not take into account student's needs will lead to a drop in learning acquisition (Gabay, 1972).

Vigil and Oller (1976) delineate two types of feedback, cognitive and affective. Within both cognitive and affective feedback there are three types, positive, negative and neutral. Cognitive feedback relays information regarding the cognitive import of the response. For example, neutral feedback can be, "I am still processing to discover its cognitive import" and negative feedback can be, "I do not understand". Affective feedback relays information on the state of the relationship For instance, positive feedback can be, "I like it". For both affective and cognitive feedback, positive feedback calls for increased output and negative feedback calls for a different kind of output and neutral feedback is somewhat ambivalent (Watzlawick, Weakland & Fisch, 1974). Negative affective feedback (eg. I do not like it) will tend to override whatever sorts of feedback that are sent on the cognitive channel (Watzlawick, Beavin & Jackson, 1967). If the feedback on the affective channel is positive the user may be apt to assume positive cognitive feedback and conversely if the affective feedback is negative he will assume negative cognitive feedback (Vigil & Oller, 1976).

Both positive affective and positive cognitive feedback have reinforcing qualities. They reinforce the student to increase his effort in the direction of additional output. Although feedback has a reinforcing quality when the students' responses are correct it also relays cognitive information about the state of the student's understanding of the material. Positive feedback relays to the student that his understanding of the material is correct as represented in his correct response. Negative feedback relays information that the student's response is incorrect. Feedback paragraphs indicating the correct answer and page numbers for further information are relaying cognitive information. This informative feedback is dynamic in the sense that it acts to effect a change in the student's response topography when his response was incorrect. The feedback paragraphs used in this present study will act as a corrective control feature after an incorrect response was given and will act as additional practice following a correct response. Its interactive nature will aid the student in transfering the material from the text.

In addition partial scoring will reinforce by giving extra marks for second and third retrys on multiple-choice items. This hopefully will strengthen the responses and also reinforce additional effort.

Summary of the Chapter

This chapter presented a review of the literature in the areas related to this study. These involved ICAT, computer-adaptive testing and partial scoring. A discussion related feedback to reinforcement and learning.

CHAPTER III

STATEMENT OF THE PROBLEM AND RESEARCH DESIGN

Statement of the Problem

The present study poses the following questions:

1. Is the number of quizzes taken and the time spent on them related to final exam performance?
2. To what extent does learning and individualization occur as a result of exposure to quizzes?
3. Is the extent of transfer affected by the nature of the material learned?
4. Are explanatory paragraphs requested by subjects related to the number of quiz repetitions and/or final performance?
5. What is the effect of motivation on quiz and criterion performance?

Quiz Taking Behavior

Cartwright and Derevensky (1977) found that the number of times quizzes were taken was only weakly correlated with final exam performance. In addition they found that the total time spent on the quizzes was not significantly related to final exam performance. The following hypotheses were designed in an attempt to replicate these findings.

Hypothesis 1

The number of times the quizzes are taken is unrelated to final exam performance.

Hypothesis 2

The amount of time spent by subjects on the quizzes is unrelated to final exam performance.

Learning

Since learning is expected to occur during the duration of the experiment, it would be expected that mean scores on a performance post-test would be higher than those on a performance pre-test. Further it is expected that ICAT will produce a greater spread of performance scores due to its more individualized approach. To test this, the following hypothesis is proposed:

Hypothesis 3

The mean and the standard deviation of criterion learning scores is larger on the post-test than on the pre-test.

Cartwright and Derevensky (1977) observed that subjects who averaged between 90-100% on the quizzes scored approximately 10% higher on the criterion test than subjects who had averaged only 80-90% on the quizzes. Hypothesis 4 was designed to predict this effect under controlled conditions.

Hypothesis 4

Students scoring significantly higher on the quizzes will tend to score significantly higher on a criterion learning test.

Scoring Reliability

The use of partial scoring with the quizzes will better predict final exam performance than the use of other scoring alternatives. Evans and Misfeldt (1974) indicated that partial scoring increased the split half reliability of the scores. However an unpublished pilot study with one quiz indicated that partial scoring did not necessarily improve prediction of criterion scores. Under more controlled conditions it may be that partial scoring will provide more precise evaluation during the quizzes of the subject's current knowledge, and hence may improve the predictability of performance on the criterion test. To test this, the following hypothesis was proposed:

Hypothesis 5

> The relationship between scores on the quizzes and the criterion test score will be significantly higher for subjects in the partial scoring treatment than for subjects in the other scoring treatments.

Transfer

To test for transfer, it was decided to create a criterion test consisting of questions which had appeared during the computer quizzes (exposed questions) and questions which had not (unexposed questions). Not all the exposed questions in the final exam were necessarily seen by the subject and the possibility of simple recall was further reduced by the time interval between taking the quizzes and the administration of the criterion test. The following hypothesis was designed:

Hypothesis 6

The number of correct responses made by the subjects on the criterion test is the same for both exposed and unexposed questions.

Cartwright and Derevensky (1977) utilized entirely exposed questions on their criterion test. The present study will attempt to determine if this transfer was not more recall than transfer due to the previous exposure to the questions.

Transfer of learning may be manifested by each student responding to the random questions and organizing them into a coherent cognitive 'gestalt' or pattern with respect to the target subject area. To test whether this integration is affected by the nature of the question it was hypothesized that transfer will be generated equally between the applied and the theoretical questions (classification of items was determined through majority consensus among three professors). The following hypothesis was generated to determine if this was so:

Hypothesis 7

The final peformance of the subjects on theoretical questions is the same as on applied questions.

Reinforcement

Traditional learning theory maintains that reinforcement can effect increased learning. For the partial scoring treatment it is expected that the application of reinforcement for additional effort, will promote more learning. This will be examined by the use of modified gain scores, from pre-test to post-test. Gain scores are used to see how much the subject learned during the year. Gain scores are computed by subtracting the subject's score on the post-test from the pre-test. Modified gain scores give a better idea of how much the subject learned during the year by viewing it as relative to how much he/she could have learned. Modified gain scores are computed by taking the gain score and dividing it by the possible amount of increase the subject could have obtained (Post - Pre/Total Possible - Pre). For example if the subject got 30 on the pre-test and 40 on the post-test his modified gain score would be, (40-30)/(50-30) which is equal to .5. The following hypothesis was proposed:

<u>Hypothesis 8</u>

> The modified gain scores on the criterion test will be highest for the students who received partial scoring treatment during the quizzes.

In addition the following hypothesis was designed in line with traditional learning theory:

<u>Hypothesis 9</u>

> The modified gain scores on the criterion test for the experimental treatment of multiple trials without partial scoring will be significantly higher than those for students in the control treatment.

This was expected due to the reinforcement given for additional effort. By asking the subjects to try again, the opportunity is given for reinforcing additional responses.

Feedback

The feedback paragraphs which are associated with each question are an optional adjunct to the quizzes. Cartwright (1971) postulated that high social learning groups may better integrate their knowledge during quizzes and therefore needed fewer quizzes to reach criterion. However that study did not use feedback paragraphs nor study their effect on integration of the material. Roper (1977) wrote that the delay of information feedback allows the student to rethink the problem and also increase the attention to the feedback. Previous studies in ICAT did not research the effect of feedback on performance. As a result of the variation among the subjects in their request for feedback paragraphs, some explorative hypotheses were generated.

Hypothesis 10

> The higher the number of explanatory paragraphs requested, the fewer quizzes required by the subject to reach criterion.

This is expected because the more knowledge a subject has on the subject material, the better he will perform on the quiz and consequently will reach the criterion level sooner. The greater the number of explanatory paragraphs requested, the greater the integration of knowledge. In addition this increase in learning may affect the subjects final exam performance. Hypothesis 11 proposes that some consistent performance on the final exam may occur.

Hypothesis 11

The higher the number of explanatory paragraphs requested by the subject the higher the final score on the criterion test.

Subjects

One hundred female and five male subjects enrolled in the post-bacculaurate one-year elementary education program were assigned to one of four sections of an introductory educational psychology course at McGill University. All four sections were taught by the traditional lecture method by various instructors and all used the same text.

Research Design

Subjects in the three groups completed five ICAT quizzes during the academic year. Each quiz contained twenty questons which were randomly selected from the bank of questions. The performance criterion was set at 80%. All subjects were given a fifty question pre-test to assess their baseline knowledge. The same fifty questions were included in the post-test to test for modified gains. Of the fifty questions twenty-five were not shown again to the students during the quizzes (unexposed questions). The other twenty-five remained in the item bank for selection and presentation during the quizzes (exposed questions).

On signing on to the first quiz, each subject was randomly assigned by the computer to one of three different groups. The random assignment helped to ensure that the effects of such variables as intelligence and typing ability would be equivalent for all treatments.

The first group (control group 1) was scored on a right-wrong (1 or 0) basis. Only one trial was given for each item. The second treatment group (experimental group 2) was scored with partial scoring. Three marks were given for the first attempt, two marks for the second attempt, and one mark for the third attempt and no marks on the fourth attempt. Thus this group was scored using a 3-2-1-0 marking system. Subjects in the third treatment group (experimental group 3) were scored on a right-wrong basis (1 or 0) as in Group 1 but were asked to continue responding until the correct choice was made (as in Group 2). KCR appeared after the correct response was made. In addition KCR could be requested and it came in the form of a feedback paragraph which indicated the correct response and page numbers in the text for further reference.

An example of a multiple-choice item followed by a feedback paragraph is presented below:

```
"The term 'associationism' is used to refer to the most
widely endorsed American learning theory because:
1. It stresses association between stimuli and responses.
2. It emphasizes that learning takes place when new
   experience is associated with a previous experience.
3. It was developed by Watson, Thorndike, Skinner, and their
   associates.
4. It was widely discussed at a convention of the American
   Psychological Association.
Please type 1, 2, 3, 4, B, or SOS
?
4
Absolutely not
Would you like an explanation?
?
yes
Now let me tell you why
The type of association stressed by the word is that between
```

a stimulus and a response -- as between the sound of a bell and food, or pushing a lever and food (pages 57-58) (option 1) Associating a new experience with an old one is a better description of field theory than associationism, since it implies insight and rearrangement of patterns of thought. Options 3 and 4 are essentially correct, but not the basic reason for the choice of the word "associationism".

Procedure

At the beginning of the academic year all students took the paper and pencil pre-test. In all of the treatments the subjects were required to complete five quizzes during the academic year. Students could repeat the quizzes as often as necessary to reach or surpass the criterion level. The computer quizzes formed 35% of the course grade. The five quizzes were developed based on chapters in the Biehler (1974) text. Students studied the relevant chapters in the textbook before doing the quizzes. Subjects in all treatment groups were given a list of dates by which they should finish and they paced themselves within these dates. The groups used the ten cathode ray tube (CRT) terminals located in the Education building of McGill University.

During the first quiz, students were asked to indicate their instructor, sex, age, academic program, and number of three credit courses previously taken in psychology. Instructions regarding their automatic group assignment were then given. This assignment was made randomly and automatically by the computer. Students remained in the same treatment throughout the study.

To review, a sixth quiz was designed to sample items from the entire text. This quiz was optional and did not influence the final course grade. The sixth quiz randomly selected twenty questions from the entire bank of items. At the end of the academic year all students were required to complete the final post-test.

Software

The Programming Language

The quizzes used in this study were coded in an author language known as CAN-VI (Cartwright & Tessler, 1975). CAN-VI is a Fortran-Based computer-assisted instruction language which permits the easy development and use of ICAT. The language is interactive and the system automatically keeps performance records for each student.

The Quizzes

The six quizzes were developed from certain chapters in the Biehler (1974) textbook. The six quizzes were:

Quiz One - Chapters 1 and 2

Quiz Two - Chapters 3 and 4

Quiz Three - Chapters 5, 6 and 7

Quiz Four - Chapters 8,9 and 10

Quiz Five - Chapters 11, 12, and 14

Quiz Six - Entire Text (optional)

The items were provided by the publisher. Along with immediate KCR for each item an elaborate feedback paragraph was associated with each item. The feedback paragraphs told the student the correct answer and why alternate answers were incorrect and contained page numbers in the text for reference.

Hardware

The hardware used in the present study included ten Volker Craig VC303 CRT's. The terminals are connected by Gandalf Data Set to an IBM 370/158 timesharing computer.

Description of Measuring Instrument

Criterion Learning Test

The criterion test consisted of fifty multiple-choice items based on the material presented in the quizzes. These were the same items as on the pre-test. Each item presented a choice of four possible answers.

Like the pre-test, the criterion learning test was a paper-and-pencil test and was administered individually to each subject.

Data Collection

The data collected consisted mainly of scores on the pre-test and post-test and the five quizzes. Students marked the answers to the pre-test and the post-test on optically read answer cards, and the results were later scored and analyzed.

The quizzes were automatically scored and the results stored on disk by the computer. The CAN-VI program which ran the quizzes recorded automatically each choice made, the number of explanatory paragraphs requested, elapsed time and the number of correct and incorrect responses.

Summary of the Chapter

This chapter presented a statement of the problem. Eleven hypotheses were designed to test the effect of certain variables on overall group performance and to investigate the effect of the three treatment groups on learning performance. A description of subjects was given and the research design and procedure were presented. In addition, a description was given of the software, hardware, measuring instruments, and procedures of data collection.

CHAPTER IV

ANALYSIS OF THE DATA

Overview of the Chapter

This chapter presents the results of analyses that were performed to test the hypotheses formulated in Chapter III. The results are presented mainly in tabular form in the order of the hypotheses.

Subjects

During the course of the year, seventeen students did not complete the course. In addition, data on pre-test scores for nine of the 105 students who completed the course were unavailable due to absence and/or late registration. As a result, data for these nine students were excluded from those analyses which required the use of pre-test scores. Table 1 indicates the number of subjects in each treatment.

Table 1

	Group			Total
	1	2	3	
all. subjects	36	28	41	105
subjects with all data complete	35	25	36	96

Results

Hypotheses 1 and 2 - Quiz Taking Behavior

Pearson product-moment correlation coefficients were calculated to test hypotheses 1 and 2 that the number of quizzes taken and the amount of time spent on the quizzes were unrelated to final exam performance. Table 2 shows the correlations concerning the hypotheses.

Table 2

Correlations for the Three Treatment Groups

Correlation	Group			all
between:	1	2	3	subjects
no. of quizzes taken and criterion test score.	-.324*	-.589***	.030	-.190**
elapsed time and criterion test score	.369*	.164	.416**	.312**

* $p \leq .05$

** $p < .01$

*** $p < .001$

For all subjects, the analysis indicates that the number of times the quizzes are taken is weakly but negatively correlated with the final performance. The analysis relating time on quizzes with the final performance shows a significant but mild correlation between the two variables. Among the correlations for the three treatment groups, there is a significant negative correlation for groups 1 and 2 between the number of quizzes taken and criterion test score. No significant relationship was found for group 3. In relating elapsed time to criterion test score, a significant positive correlation was found for groups 1 and 3.

Table 3

Correlations

Correlation between:	
no. of quizzes taken and elapsed time	.535*
no. of psych. courses taken and criterion test score	-.167

*p<.001

As might be expected, the results in Table 3 indicate a highly significant positive correlation between the number of times the quizzes were taken and the total elapsed time. No significant correlation exists between the number of previous psychology courses students had taken and the criterion test score.

Table 4 presents the means and standard deviations of the number of quizzes taken, number of explanations requested, elapsed time, average time per quiz, and average quiz score for the three treatments.

Table 4

Means and Standard Deviations for the Three Treatments

		Group			all
		1	2	3	subjects
No. of Quizzes taken:	Mean	18.00	15.00	19.70	17.90
	s.d	7.83	5.95	9.09	8.06
No. of Explanations	Mean	171.00	154.00	162.00	163.00
requested:	s.d.	96.80	81.60	99.50	93.50
Total Elapsed time	Mean	516.90	543.60	586.00	551.00
(in minutes)	s.d.	211.00	251.90	260.70	240.70
Average Time per	Mean	28.70	36.20	29.70	30.80
Quiz (in minutes)					
Average Quiz Score	Mean	86.60	90.80	87.80	88.20
	s.d.	9.16	3.80	4.35	4.09

On the number of quizzes taken, group 3 has the highest mean and standard deviation followed by groups 1 and 2. On the number of explanations requested, group 1 has the highest followed by group 3 and 2. On elapsed time, group 3 has the highest mean followed by groups 2 and 1. Dividing the total elapsed time by the mean number of quizzes gives the average amount of time per quiz. It will be noted that group 2 took the most time per quiz followed by groups 3 and 1 respectively. On average quiz score, group 2 has the highest followed by groups 3 and 1. Analyses of variance were performed on each of the variables in table 4 but only the average quiz score variable showed significant differences among the group means (F=10.03, df=2, p<.001).

Hypothesis 3 - Learning

Hypothesis 3 predicted that the post-test would show a higher mean and a larger standard deviation. The means and standard deviations are presented in Table 5.

Table 5

Means and Standard Deviations of

Criterion Learning Scores for the Pre-Test and Post-Test

Variable	N	Mean	s.d.
Pre-Test	96	21.1	4.98
Post-Test	96	35.2	5.92

A t-test between pre and post-test was found to be significant (t=21.8, df=95, p<.001) confirming the hypothesis that students had learned during the course of the year.

Hypothesis 4 - Achievement

For this hypothesis, the average quiz score for the five quizzes was correlated with the post-test score. These data were divided into two groups, a low quiz score group which averaged between 80-89% on the quizzes, and a high quiz score group which averaged between 90-100%.

Table 6

Means and Standard Deviations
of the High and Low Quiz Score Groups

Group	N	Mean	s.d.
Low quiz score	72	67.8	5.93
High quiz score	33	76.2	4.24

A significant difference was found between means in the direction of the 90-100 group (Welch t'=-4.123, df=103, p<.001) Another analysis was performed to determine whether subjects who did better on the quizzes did better on the post-test. A significant correlation was found (r=.312, p<.01).

Hypothesis 5 - Scoring Predictability

The results for the three treatment groups are presented in Table 7.

Table 7

Correlations for the Three Treatment Groups

Correlation	Group		
between:	1	2	3
average quiz score and criterion test score	.155	.590**	.242

** p<.001

The table shows a highly significant correlation between the average quiz score and the criterion test score for the partial scoring treatment, but not for groups 1 and

3.

Table 8 presents the data divided into exposed and unexposed post-test scores.

Table 8

Correlations for the Three Treatment Groups

Correlation	Group			all
between:	1	2	3	subjects,
average quiz score and exposed post-test score	.282	.609**	.348*	.381**
average quiz score and unexposed post-test score	-.035	.340	.101	.146

* p<.05

** p<.001

It should be noted that the correlations are larger between average quiz score and exposed questions than for average quiz score and unexposed questions.

Hypotheses 6 and 7 - Item Characteristics

Hypotheses 6 and 7 predicted that the number of correct responses made by subjects on the criterion test would be the same for both theoretical versus applied items and for exposed versus unexposed questions. Table 9 presents the results for these variables.

Table 9

Means and Standard Deviations

post-test scores for:	N	Mean%	s.d.
exposed questions	105	81.6	3.44
unexposed questions	105	59.2	3.38
theoretical questions	105	70.9	4.06
applied questions	105	64.4	2.30

Subjects had more correct responses for the exposed questions. A t-test was performed and the difference was found to be significant (t=15.8, df=104, p<.001) This suggests that students benefited on the post-test from experiencing items previously during the quizzes. In addition subjects had significantly more correct responses for the theoretical questions (t=5.69, df=104, p<.001)

Hypotheses 8 and 9 - Reinforcement

For hypotheses 8 and 9 it was predicted that the partial scoring treatment would be most reinforcing and that the two experimental treatments would be more reinforcing than the control treatment.

Table 10 presents the means and standard deviations of pre and post-test data.

Table 10

Means and Standard Deviations for the Three Treatments

		Group			all
		1	2	3	subjects
Pre-Test	Mean	21.50	20.60	20.90	21.05
	s.d.	4.00	5.88	6.98	5.77
Post-Test	Mean	35.20	35.90	34.70	35.20
	s.d.	4.00	5.88	6.98	5.74

Modified gain scores were computed and are presented in Table 11.

Table 11

Modified Gain Scores for the Criterion Test

Treatment	N	Mean	Variance
1	35	.484	.018
2	25	.522	.029
3	36	.460	.062
Total	96	.485	.037

An F-test was performed on gain scores for the three treatments but no significant difference was found (F=.76, p=.47). Due to the inherent weakness in modified gain scores, the analysis was repeated using analysis of covariance with treatments as the independent variable, post-test scores as the dependent variable and pre-test scores as the covariate. The results indicated no significant differences. The results do not support the hypotheses. The experimental conditions did not result in higher scores on the criterion learning test. Another analysis of covariance was performed using treatments as the independent variable, post-test as the dependent variable and age of student, instructor, number of psychology courses taken and pre-test scores as covariates. The results also indicated no significant differences.

Hypotheses 10 and 11 - Feedback

Hypotheses 10 and 11 predicted that the more explanatory paragraphs requested, the fewer quizzes the subject would take and the higher the score on the criterion test. Table 12 presents the correlations between the two variables in each hypothesis.

Table 12

Correlations for the Three Treatment Groups

Correlation	Group			all
between:	1	2	3	subjects
no. of explanations requested and number of quizzes taken	.724**	.549*	.559**	.604**
no. of explanations and criterion test score	-.189	-.248	-.162	-.186

* p<.01

** p<.001

The results indicate that the number of explanations requested is related to an increase in the number of quizzes taken. No significant differences were found between the correlations for group 2 and 1, or group 3 and 1. There was no significant relationship found between the number of explanations requested and the criterion test scores.

Summary of the Chapter

This chapter presented the results of the statistical analyses performed on the data. The results suggest that the subjects who take more quizzes usually do not do as well on the final test and that more time spent on quizzes resulted in higher performance. Higher performance on the quizzes resulted in higher performance on the criterion test. Partial scoring appeared to be a better predictor of final exam performance.

In addition the results suggested that subjects learned more about theoretical information than applied information. Also the exposure to the questions resulted in better performance. Reinforcement due to treatment did not significantly affect performance.

With regard to feedback the feedback paragraphs did not reduce the number of quizzes taken, nor did the number of explanations requested affect performance on the criterion test.

CHAPTER V

DISCUSSION

Overview of the Chapter

This chapter evaluates the results presented in Chapter IV. Statements regarding the support or the non-support of the hypotheses are presented. Implications for education are discussed and suggestions for further research are given.

Learning and ICAT

Quiz Taking Behavior

The results do not support hypotheses 1 and 2. The negative correlation between number of times the quizzes were taken and the final exam score suggest that the higher the number of quizzes taken, the lower the criterion performance scores. For hypothesis 2, the results suggest that the more time the subjects spent on the quizzes, the higher their final performance. These results are contrary to the findings of Cartwright and Derevensky (1977) who found that the number of times the quizzes were taken was only weakly, but positively, correlated with final exam performance. In contrast, table 2 shows a significant but weak negative correlation for all subjects.

In addition, Cartwright and Derevensky (1977) found that the total time spent on quizzes was not significantly related to final exam performance, while the results of the present study suggest that the variables are significantly related. In addition in table 2, the control treatment (group 1), which typifies the methodology of Cartwright and Derevensky's study, also shows a significant negative correlation between the number of times quizzes were taken and final criterion performance, as well as a significant positive correlation between elapsed time and final performance.

In view of the negative correlation between quiz taking behavior and post-test scores, one possible explanation which could account for why students continued taking quizzes which did not appear to help their performance on the post-test, is that subjects may have used the quizzes in different ways. For example, subjects who knew their material did not need to repeat the quizzes as often and probably studied the text before repeating the quiz. On the other hand, certain subjects may have paid too much attention to the quizzes instead of studying the text as a way of learning the subject material. To these students the quizzes became a competitive way of learning instead of a way of obtaining feedback on how well they studied the text. However, subjects who spent more time on each of the quizzes may have integrated their knowledge more. Support for this lies in the significant positive correlation between elapsed time and criterion test score.

Looking at the group differences in table 4 for the number of times quizzes were taken, the least number of quizzes appears taken by subjects in the partial scoring treatment (group 2). This might be expected since subjects in the partial scoring treatment might find it easier to achieve the mastery score of 80% since they are constantly benefitting from fractions of marks on each item. Subjects in the partial scoring treatment got significantly higher quiz marks, the largest gain being in the order of 4%. The partial scoring treatment also appeared to take more time than the control group 1, but this difference was found to be not significant though the direction was as predicted. Logically, multiple responding for the item retrys would be expected to increase elapsed time on the quizzes. This difference however, disappeared (see table 10) in a comparison of group means on the criterion test. While the reasons for this are not immediately clear, it could be either that the original gain on the quizzes was of short term duration, or that the control group compensated for the lower quiz scores by additional study of the text.

Learning

The results in table 5, support hypothesis 3 that the mean and standard deviation is higher on the post-test than on the pre-test. For hypothesis 4, the results are supportive that students as a group who do better on the quizzes do better on the final criterion test (table 6). This was also the finding of Cartwright and Derevensky (1977) who found that subjects who averaged between 90-100% on the quizzes scored approximately 10% higher than subjects who averaged between 80-90% on the quizzes.

In addition the overall correlation between average quiz score and final performance (table 7) is significant, and supports the assertion of Cartwright and Derevensky (1977) that the greater the criterion level of mastery for the quizzes, the greater the performance on the final test. But the question still remains: does the level of mastery for the quizzes affect the final score or is it simply the effect of some common variable such as intelligence, which correlates with both quiz performance and criterion performance?

Scoring Predictability

The results in table 7 support the hypothesis that the relationship between the scoring procedure for the quizzes and the final exam score is highest for the partial scoring treatment. This supports the concept of partial scoring as a more precise evaluation of the subject's knowledge. Although Evans and Misfeldt (1974) indicated that partial scoring increases the split-half reliability of the test, the results of the present study suggests that scores for the partial scoring treatment correlate more highly with the criterion test score.

It is interesting to see the correlation between average quiz score, and exposed and unexposed questions, where a weaker relationship for the unexposed items than for the exposed was predicted. In table 8, the partial scoring treatment has a higher correlation between average quiz score and the score on exposed questions on the post-test. This is expected because exposed questions are those exposed to the students in the quizzes and consequently a stronger relationship would be logical due to simple recall.

Item Characteristics

The results shown in Table 9 do not support hypotheses 6 and 7. Subjects had more correct responses for both the exposed questions and for the theoretical questions than for the unexposed and the applied questions, respectively. It seems that due to previous exposure, subjects knew more of the answers to the exposed questions. Students did significantly better on the theoretical questions on the post-test while on the pre-test students did slightly better on the applied questions. This is probably because students coming into the course may have had more ideas about applying educational principles than knowledge of content or theoretical issues. However it is gratifying to note that during the course of the year students improved their theoretical knowledge.

To account for the gain in theoretical over applied knowledge, one could theorize that theoretical principles have a greater transfer capability than applied information, since theoretical principles are less specific and hence are more generalizable. Theoretical principles can build more associations with other items in memory. Rather than emphasizing the characteristics of long term and short term memory, Craik and Lockhart (1972) stress the amount of processing the item has received as the major determinant of the characteristics of memory. Theoretical knowledge may undergo more processing and elaboration and is therefore better represented in the memory store, and hence easier to retrieve.

Reinforcement

The results in Table 11 do not support hypotheses 8 and 9. The added reinforcement for the experimental conditions especially for the partial scoring condition did not effect an increase in performance on the criterion test. Although, the highest gain score appears to be for the partial scoring treatment, and is in the predicted direction it does not differ significantly from the others. It seems that the quizzes were not the only sources of reinforcement. For these hypotheses to be better tested it would be necessary to control other sources of reinforcement, leaving the scoring treatments as the only sources of external reinforcement. This would be difficult because one can not control all the various sources of reinforcement in an educational environment. Table 10 shows the means for the groups and indicates that the added reinforcement for the experimental conditions did little to effect better performance on the post-test.

Feedback

The results in table 12 do not support hypotheses 10 and 11. The feedback paragraphs did not appear to add to the subjects' body of knowledge in a significant way to effect an increase in the final exam performance. Rather, the more quizzes the subject took the more explanatory paragraphs he or she requested. Looking at the correlations between number of explanations requested and final criterion performance, it appears none were found to be significant, suggesting that the explanatory paragraphs added little to the overall knowledge. As mentioned earlier many subjects probably did not study the text and consequently the explanatory paragraphs did not encourage them to study the text further. Instead, many subjects may have continued retaking the quizzes, possibly hoping to assimilate solely the target knowledge through the items and the explanatory paragraphs.

Implications for Education and Suggestions for Further Research

The conclusions of this study are limited by the scope of the research conducted. This area of research was both an experiment and part of an on-going course in an educational institution, with all the limitations that it implies. Nevertheless, as indicated by the increase in scores on the post-test, learning was achieved. The question that remains is what role did ICAT play in that educational experience? As noted earlier in chapter 2, ICAT with its feedback paragraphs was intended to serve as an effective teacher as well as an effective tester. It appears that its testing role has been refined with the addition of partial scoring, making ICAT a more sophisticated evaluation tool.

As for ICAT's role as a learning catalyst, it seems that the feedback paragraphs added little to the assimilation of the target knowledge. As mentioned earlier, students may have taken advantage of ICAT by using it as a way of learning the subject material. It is possible that the feedback paragraphs plus the lack of an upper limit on the number of times quizzes could be repeated projected the image that the computer quizzes are an easy way of learning the subject material and contributed to reduced study of the text. For this reason it is recommended that future research incorporate a limit for the number of times quizzes may be repeated. ICAT with its feedback paragraphs is mainly a testing tool but appears to have the capability of allowing the student to learn during testing.

Further research might examine what effect limiting the extent of feedback has on knowledge acquisition, either by limiting the amount of information that is in the feedback paragraph or by limiting the number of requests for feedback paragraphs in a testing session. A different kind of balance between the testing and the teaching role may need to be defined before ICAT can achieve its true potential.

Further research might concentrate as well on how ICAT can reinforce learning and what role it plays as an incentive for learning in the overall academic experience. This study has indicated that the reinforcement given for retrys did not significantly effect an increase in performance on the criterion test. It would be interesting to study other ways in which ICAT can reinforce learning.

This study has shown that the nature of the subject material affects performance. Previous exposure to items had critical effects on final performance. Theoretical items were shown to have better transfer capability. Further research might examine this result in more detail under more controlled conditions.

Conclusion

This study has attempted to offer an alternative approach to interactive computer-assisted testing. It has also presented variations in testing and scoring procedures. This study investigated computer usage in both its learning and testing roles. Partial scoring was found to be a better predictor of final performance. The results of the study suggest that students' requests for feedback paragraphs did not lead to an increase in criterion performance. In addition experimental treatments in applying reinforcement for item retrys did not effect an increase in final performance.

This study has shown that quiz taking frequency is negatively related to final performance and that elapsed time on quizzes is significantly related to final performance. Exposed and theoretical items were found to effect greater recall than unexposed or applied items. This study confirms a previous finding that the higher the criterion mastery level on the quizzes, the higher the post-test score.

It is suggested that further research might show how a better balance might be achieved in the role of ICAT as both teacher and tester. In addition further research is recommended to find out how ICAT can play a greater role in the acquisition of knowledge.

References

Anderson, R. C., Kulhavy, R. W. and Andre, T. Feedback procedures in programmed instruction. *Journal of Educational Psychology*, 1971, *62*, 148-156.

Bayroff, A. G. Feasibility of a programmed testing machine. Washington: U.S. Army Personnel Research Office, 1964.

Betz, N. E. & Weiss, E. J. *Psychological effects of immediate knowledge of results and adaptive testing.* Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1976.

Biehler, R. F. *Psychology Applied to Teaching*, 2nd edition, Boston: Houghton Mifflin, 1971.

Biehler, R. F. *Psychology Applied to Teaching*. Boston: Houghton Mifflin, 1974.

Buss, A. H., Braden, W., Orgel, A., & Buss, E. Acquisition and extinction with different verbal reinforcement combination. *Journal of Experimental Psychology*, 1956, *52*, 288-295.

Cartwright, G. F. Social factors in computer-assisted testing. *Proceedings of the McGill University Conference on University Teaching and Learning*. Montreal: McGill University, 1971.

Cartwright, G. F. *Social, personality, and attitudinal dimensions of individual learning with computer-assisted group instruction*. Unpublished doctoral dissertation. Edmonton: The University of Alberta, 1973.

Cartwright, G. F. A promising innovation: computer-assisted test construction. Learning and Development, 1975, 6, 1-3

Cartwright, G. F. & Derevensky, J. L. An attitudinal study of computer-assisted testing as a learning method. Psychology in the Schools, 1976, 13, 317-321.

Cartwright, G. F. & Derevensky, J. L. The development of computer-assisted testing as an adjunct to traditional instructional processes in educational psychology. Paper Presented at the Canadian Educational Researchers Association. Fredericton, New Brunswick, 1977.

Cartwright, G. F. & Tessler F. Course author's guide to computer-assisted instruction using CAN VI. Montreal: Centre for Learning and Development, McGill University, 1975.

Cleary, T. A., Linn, R. L. & Rock, D. A. An explanatory study of programmed tests. Educational and Psychological Measurement, 1968, 28, 345-360.

Cohen, P. The effect of discussion, individual response and feedback on learning and attitudes of individuals in a group computer-assisted instruction setting. Unpublished M.A. Thesis. Montreal: McGill University, 1975.

Craik, F. I. M., & Lockhart, R. S. Levels of processing: A framework for memory research. Journal of Verbal Learning and Verbal Behavior, 1972, 11, 671-684.

Derevensky, J. L. & Cartwright, G. F. The use of computer-assisted testing in an introductory course in educational psychology. In Hills, P. and Gilbert, J. (Eds.) Aspects of Educational Technology XI: The spread of Educational Technology London: Kogan Page, 1977, 454-457.

Dewitt, L. J. & Weiss, D. J. A computer software system for adaptive ability measurement. Minneapolis: University of Minnesota, 1974.

Evans, R. M. & Misfeldt, K. Effect of self-scoring procedures on test reliability. Perceptual and Motor Skills, 1974, 38, 1248.

Evans, R. M. & Surkan A.J. Computer-delivered testing. CATC Digest, 1977a, 2, 3.

Evans, R. M. & Surkan, A. J. Effect of computer-delivered testing on achievement in a mastery learning course of study with partial scoring and variable pacing. Proceedings of Eighth Conference on Computers in the Undergraduate Curricula, Michigan State University, 1977b.

Feldhussen, J. F. & Birt, A. A study of nine methods of presentation of programmed learning material. Journal of Educational Research, 1962, 55, 461-466.

Ferguson, R. L. & Hsu, T. The application of item generators for individualized mathematics testing and instruction. Pittsburgh: University of Pittsburgh, Learning Research and Development Center, 1971.

Franklin, S. & Marasco, J. Interactive Computer-Based Testing. Journal of College Science Teaching, 1976, 13, 15-20.

Gabay, T. V. Peculiarities of Acquisition under the Realisation of Behavioristic Principles in Programmed Instruction. Unpublished P.H.D. Dissertation, Moscow: The University of Moscow, 1972.

Gilman, D.A. Comparisons of several feedback methods for correcting errors by computer assisted instruction. Journal of Educational Psychology, 1970, 60, 503-508.

Hansen, D. N., Johnson, B. F., Fagan, R. L., Tam, P. & Dick, W. Computer-based adaptive testing models for the Air-Force Technical Environment Phase I: Development of a computerized measurement system for Air Force Technical Training. Brooks Air Force Base, Texas: Air Force Human Resources Laboratory, 1974.

Holland, J. G., & Skinner, B. F. The analysis of behavior. New York: McGraw-Hill, 1961.

Kulhavy, R, W. Feedback in written instruction. Review of Educational Research, 1977, 47, 211-232.

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

Lippey, G. Observations, CATC Digest, 1977, 2, 3.

Lord, F. M. Some test theory for tailored testing. In W.H. Holtzman (Ed.), Computer-assisted instruction, testing and guidance. New York: Harper and Row, 1970.

Lord, F. M. Robbins Munro procedures for tailored testing. Educational and Psychological Measurement, 1971a, 31, 3-31.

Lord, F. M. The self-scoring flexilevel test. Journal of Educational Measurement, 1971b, 8, 147-151.

Lord, F. M. Tailored testing, an application of stochastic approximation. Journal of the American Statistical Association, 1971c, 66, 707-711.

Lord, F. M. Theoretical study of the measurement effectiveness of flexilevel tests. Educational and Psychological Measurement, 1971d, 31, 805-813.

Lord, F. M. A theoretical study of two-stage testing. Psychometrika, 1971e, 36, 227-241.

Lublin, S. C. Reinforcement schedules, scholastic aptitude, autonomy need and achievement in a programmed instruction course. Journal of Educational Psychology, 1965, 56, 295-302.

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

Roper, W. J. Feedback in C.A.I. Programmed Learning and Educational Technology, 14, 1977.

Sassenrath, J. M., & Yonge, G. D. Effects of delayed information feedback cues in learning and retention. Journal of Educational Psychology, 1969, 60. 174-177.

Skinner, B.F. Teaching Machines. Scientific American, 1961, 90-102.

Sturges, P.T. Verbal delay and retention: Effect of information in feedback and tests. Journal of Educational Psychology, 1972, 63, 32-43

Sturges, P.T. Delay of Informative Feedback and Computer Managed Instruction. Paper Presented at American Educational Research Association, New York, 1977.

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

Tait, K., Hartley, J. R. & Anderson, R. C. Feedback procedures in computer- assisted arithmetic instruction. British Journal of Educational Psychology, 1973, 43, 161-171.

Talyzina, N. F.. Psychological basis of programmed instruction. Instructional Science, 1973, 2, 243-280.

Vigil, N. A., Oller, J. W. Rule Fossilation: A Tentative Model. Language Learning, 26, 1976, 281-295.

Watzlawick, P., Beavin, J. H. & Jackson, D. D. Pragmatics of Human Communication: A study of Interactional Patterns, Pathologies and Paradoxes. New York: W.W. Norton, 1967.

Watzlawick, P., Weakland, J. W. & Fisch, R. Change: Principles of Problem Formation and Problem Resolution. New York: W.W. Norton, 1974.

Weiss, D. J. The stratified adaptive computerized ability test. Minneapolis: University of Minnesota, 1973.

Weiss, D. J. & Betz, N. D. *Ability measurement: conventional or adaptive*? Minneapolis: University of Minnesota, 1973.

Weiss, D. J. *Computerized ability testing, 1972-1975*. Minneapolis: University of Minnesota, 1976.