THE EFFECTS OF COMPUTER-BASED TESTS ON THE ACHIEVEMENT, ANXIETY AND ATTITUDES OF GRADE 10 SCIENCE STUDENTS

by

CHRISTINE HUI LI CHIN

B.Sc. (Hons), National University of Singapore, 1982
Dip.-in-Ed., National University of Singapore, 1983
M.Sc., University of Toronto, 1986

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF ARTS

in

THE FACULTY OF GRADUATE STUDIES

(Department of Mathematics and Science Education)

We accept this thesis as conforming
to the required standard

Dr. J. Stuart Donn)

(Dr. Marv Westrom)

(Dr. Robert F. Conry)

THE UNIVERSITY OF BRITISH COLUMBIA
June 1990

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of Mathematics & Science Education

The University of British Columbia
Vancouver, Canada

Date July 11, 1990

Abstract

The purpose of this study was to compare the achievement and test anxiety level of students taking a conventional paper-and-pencil science test comprising multiple-choice questions, and a computer-based version of the same test. The study assessed the equivalence of the computer-based and paper-and-pencil tests in terms of achievement scores and item characteristics, explored the relationship between computer anxiety and previous computer experience, and investigated the affective impact of computerized testing on the students.

A 2 X 2 (mode of test administration by gender) factorial design was used. A sample of 54 male and 51 female Grade 10 students participated in the study. Subjects were blocked by gender and their scores on a previous school-based science exam. They were then randomly assigned to take either the computer-based test or the paper-and-pencil test, both versions of which were identical in length, item content and sequence. Three days before the test, all students were given the "Attitude questionnaire" which included pre-measures of test and computer anxiety. Immediately after taking the test, students in the computer-based group completed the "Survey of attitudes towards testing by computers" questionnaire which assessed their previous computer experience, their test anxiety and computer anxiety level while taking the test, and their reactions towards computer-based testing. Students in the paper-and-pencil test group answered the "Survey of attitudes towards testing" questionnaire which measured their test anxiety level while they were taking the paper-and-pencil test.

The results indicate that the mean achievement score on the science test was significantly higher for the group taking the computer-based test. No significant difference in mean scores between sexes was observed; there was also no interaction effect between mode of test administration and gender. The test anxiety level was not significantly different between the groups taking the two versions of the test. A significant relationship existed between students' prior computer experience and their computer anxiety before taking the test.

However, there was no significant relationship between previous computer experience and the computer anxiety evoked as a result of taking the test on the computer. Hence, the change in computer anxiety due to taking the test was not explained by computer experience. Of the students who took the computer-based test, 71.2 % said that if given a choice, they would prefer to take the test on a computer. Students indicated that they found the test easier, more convenient to answer because they did not have to write, erase mistakes or fill in bubbles on a scannable sheet, and faster to take when compared to a paper-and-pencil test. Negative responses to the computer-based test included the difficulty involved in reviewing and changing answers, having to type and use a keyboard, fear of the computer making mistakes, and a feeling of uneasiness because the medium of test presentation was unconventional. Students taking the computer-based test were more willing to guess on an item, and tended to avoid the option "I don't know."

It is concluded that the computer-based and the paper-and-pencil tests were not equivalent in terms of achievement scores. Modifications in the way test items are presented on a computer-based test may change the strategies with which students approach the items. Extraneous variables incidental to the computer administration such as the inclination to guess on a question, the ease of getting cues from other questions, differences in test-taking flexibility, familiarity with computers, and attitudes towards computers may change the test-taking behaviour to the extent that a student's performance on a computer-based test and paper-and-pencil test may not be the same. Also, if the tasks involved in taking a test on a computer are kept simple enough, prior computer experience has little impact on the anxiety evoked in a student taking the test, and even test-takers with minimal computer experience will not be disadvantaged by having to use an unfamiliar machine.

Table of contents

List of Tables

List of Figures

Acknowledgements

I wish to thank my supervisor Dr. Stuart Donn, for his help, advice, and understanding during the writing of this thesis. I am especially grateful to him for critically reading, editing, and correcting the drafts of the thesis, as well as providing many valuable and constructive suggestions. His time spent on the tedious task of proof-reading the initial drafts is very much appreciated. I am also particularly indebted to Dr. Robert Conry for being so generous with his time and expertise throughout the course of this research study. I am extremely appreciative of his guidance in the development of the instruments used in this thesis, his counsel in the design and implementation of the study, his assistance with the statistical analyses and interpretation of results, his many interesting ideas, and his consistent and good-natured support. His immeasurable help in shaping the form and content of this thesis cannot be adequately acknowledged; I hope my heartfelt thanks will suffice. I would also like to thank Dr. Marv Westrom for his ideas in the design of the computer-based test used in this study.

My sincere thanks are also extended to Thomas Garcia, a fellow graduate student in my department, for his capable and kind help with programming the computer-based test in Hypercard. And last, but not least, I would like to express my gratitude to the staff and students of Alpha secondary school in Burnaby, B.C. for their cooperation and participation in this study.

Chapter 1 INTRODUCTION

As microcomputers become increasingly common in educational settings, computer administered testing with its advantage of speed, flexibility and efficiency is coming into vogue. In comparison to paper-and-pencil testing, computerized testing is a novel experience to most students. Although guidelines exist for the construction of computerized tests (Mizokawa & Hamlin, 1984), there is little empirical evidence assessing the impact of computerized testing on student performance. Many research questions remain unanswered. For example, will students with little or no previous computer experience be disadvantaged when taking a computerized test? Does gender affect test-taking on the computer? What are examinees' overall reactions to computerized testing? An important concern related to the use of computers in testing is whether results obtained from computer-based tests are equivalent to those from conventional paper-and-pencil tests in terms of achievement scores. This issue, as well as the effects of computer testing on the test anxiety and attitudes of test takers, is the major focus of the research study.

1.1 Background of the Problem

In the context of testing, "validity" means the degree to which a test actually measures what it is intended to measure. The purpose of a computerized test is to assess the examinee's knowledge and competence in the area being tested, not the examinee's computer familiarity or literacy. Effective use of a computer in a testing context however, demands that the examinee be able both to identify the correct answer to a problem and to properly communicate the answer via the computer. Accordingly, the primary concern regarding the administration of tests on a computer is whether irrelevant extraneous variables incidental to the computer administration, such as previous experience or familiarity with computers, or attitudes towards computers, either facilitate or inhibit the examinee's performance on the computerized test. These computer-linked factors may change the nature of the test-taking task

to the extent that the computerized and the conventional paper-and-pencil versions of a test do not measure the same construct. In that case, the validity of the computerized tests may be threatened if performance is associated with level of knowledge, experience with, and attitudes towards computers. Validity, then could not be generalized from the computerized to the conventional version of a test, since examinees would not attain the same score on the computer as if tested conventionally.

There is a possibility that lack of computer familiarity and anxiety on the part of some test-takers may unfairly handicap their performance on the computerized test (Lee, 1986; Llabre, Clements, Fitzhugh, Lancelotta, Mazzagatti, & Quinones, 1987). Changes in test format caused by computerization could sufficiently alter items so that different skills are required than are needed for the conventional version. People accustomed to working with computers could well have an advantage when taking a computer-based test, when compared to novices whose normal test anxiety is compounded when they are confronted with an unfamiliar machine. To the extent that achievement is influenced by such variables as experience, knowledge, or attitudes towards computers, the validity of test scores is a measurement concern. This potential problem is compounded by the fact that variables such as computer experience may be related to gender. There is evidence suggesting that males and females have different experience with, knowledge of, and attitudes towards computers (Chen, 1986; Levin & Gordon, 1989; Popovich, Hyde, & Zakrajsek, 1987), with males having greater exposure to computers and holding more positive attitudes.

A number of studies have compared the effects of computer-based tests with conventional paper-and-pencil tests on examinees' performance and anxiety (e.g. Lee, 1986; Llabre *et al*, 1987; Plumly & Ray, 1989; Ward, Hooper, & Hannafin, 1989; Wise, Barnes, Harvey, & Plake, 1989). Almost all of these studies have focused on college students or adult populations. Relatively little research has investigated the effects of mode of test

administration with elementary or high school students. The results of the aforementioned studies are not consistent, with some indicating equivalence of mean achievement scores between the two test versions (e.g. Plumly & Ray, 1989), and others showing significantly lower scores on the computer-based test (e.g. Llabre *et al*, 1987). The inconsistencies in findings are possibly due to differences in test domains and designs, failure to control for test-taking flexibility, anxiety, familiarity with computers or other variables. These findings however, stress the need to consider human factors concerns when using computers in testing. Different persons with the same ability, as measured by their scores on an achievement test, may not perform equally well on a computerized test. Individual differences variables such as computer anxiety and computer experience may have a confounding effect on computer-administered test scores. Computerized testing may discriminate against those who have not worked with computers prior to testing. Unequal access to computers in schools could also perpetuate disparities between subgroups if computerized tests are used to measure achievement. Accordingly, an important area for investigation prior to implementation of computer-based testing is whether the computerized test discriminates against those who have less computer experience or who have computer anxiety. Other practical issues to be considered when a test is modified by computer-administration are the reliability of the test and the comparability of scores in each version. This equivalence of scores in the two versions must also be established to provide evidence that both versions are measuring individual differences in a similar manner.

The need to ascertain the equivalence between a computer-based and conventional paper-and-pencil test, a consideration of the affective impact of computerized testing procedures on the student, and the lack of findings in this area at the high school level resulted in several research questions which gave rise to this study. It is essential that the effects of computer-based testing be thoroughly understood, so that the benefits of this mode of testing can be fully realized while maintaining the quality of the test results.

1.2 Statement of the Research Problem

The purpose of this study is to compare the relative performance and test anxiety level of a computer-based science achievement test (CBT) with a conventional paper-and-pencil version of the same test (PPT), for Grade 10 students in a classroom setting. The study assesses the differential effect, if any, of gender on test performance in both the computer-based and conventional paper-and-pencil tests. It will also investigate examinees' reactions to computerized testing and determine if the amount of previous experience with computers affect the computer anxiety and achievement scores of students tested via the computer. In the study, the test taking flexibility is made as similar as possible for each presentation medium, and the effects of medium of presentation are considered both at the level of total test score and at the level of individual items.

Specifically, the following research questions are addressed:

1. Does the format of the test (paper-and-pencil vs. computer-based) result in a difference in mean achievement scores of Grade 10 science students?

2. Are there gender differences in mean achievement scores when males and females are tested by the two formats?

3. Is there an interaction effect between mode of test administration and gender?

4. Are the computer-based and paper-and-pencil test versions equivalent in terms of variances, shapes of score distributions, and item characteristics?

5. Is there a difference in test anxiety level between Grade 10 science students taking the paper-and-pencil and computer-based test?

6. What is the relationship between the amount of previous computer experience and the computer anxiety level of students taking the computer-based test?

1.3 Research Hypotheses

The hypotheses corresponding to the above research questions are as follows:

1. There is no difference in mean achievement scores of Grade 10 science students tested by the paper-and-pencil test and computer-based test.

2. There is no difference in mean achievement scores between males and females tested by the two different formats.

3. There is no interaction between mode of test administration interaction and gender, i.e. the effects of the computer-based test and paper-and-pencil test will be the same for both males and females.

4. The computer-based and paper-and-pencil test versions are equivalent in terms of variances, shapes of score distributions, and item characteristics.

5. There is a difference in test anxiety level between students tested by the paper-and-pencil and computer-based tests, with students taking the computer-based test reporting a higher level of anxiety.

6. Subjects with more computer experience will report lower levels of computer anxiety.

1.4 Operational definitions of terms

The terms relevant to this research study are defined below.

1. *Conventional paper-and-pencil test*

   This refers to presenting test questions and accepting responses from examinees by the traditional paper-and-pencil method. The questions are presented on paper, and test-takers respond by writing a, b, c, d, or e into a box at the bottom of each question frame. The test is described in detail in Chapter 3.

2. *Computer-based test / computerized test*

   These two terms are used synonymously. They refer to using the computer to administer a test that is identical in length, item content and sequence to the conventional paper-and-pencil test. Responses from examinees are keyed into the computer by the test taker. The test is described in detail in Chapter 3.

3. *Science achievement*

   This is measured by the number of correct responses on the test.

4. *Test equivalence*

   This is assessed in terms of the observed mean scores, variances and shapes of score distributions of the computer-based test and the paper-and-pencil test.

5. *Test anxiety*

   This is measured by the "Attitudes towards testing" instrument designed for this study to assess examinees' test anxiety while taking the test.

6. *Computer anxiety*

This is measured by the "Attitudes towards computer-based test" instrument designed for this study to assess examinees' computer anxiety while taking the computer-based test.

7. *Power test*

In a power test, each examinee depending on his or her ability can correctly answer only a certain number of items even without time limits. Often, time limits are generous enough to ensure that each examinee can attempt each item (Allen & Yen, 1979).

1.5 Significance of the study

The importance of research in this area was alluded to in the early part of the chapter. This study compares the equivalence of a conventional paper-and-pencil science test and a computer-based version of the same test in a classroom setting. In the light of the general problem of whether the computerized version a test provides results that are comparable to that of a traditional paper-and-pencil test, this study will help to answer some of these pertinent questions. In addition, it seeks to determine the problems that test-takers encounter on a computer-based test. The results obtained would shed light on some of the factors indigenous to computer administration, but irrelevant to the purposes of the test, which may alter test performance. These findings may well have implications for test administrators and designers of computerized testing software, who should be cognizant of the effects of modifications in the method of presenting stimulus material, differences in the task required of the respondents introduced as a result of computer presentation, and differences in the format for recording responses.

The study also assesses the reactions of students taking the computer-based test. This issue merits consideration, as students' attitudes towards computerized testing may enhance or impede the testing process. It is essential therefore, for us to be aware of the

accompanying affective impact of computer-based tests, which is still not well understood. The attitudes that students have are an important factor in whether computer-based test programmes become an effective part of the curriculum of a school system. Awareness of what attitudes students hold can assist educators in curriculum planning, in evaluating the role of microcomputers in computer-based testing, and in the future local development of a curriculum which incorporates the use of computers.

This study will also contribute to the existing knowledge about the relative contributions of previous computer-related experience on computerized testing performance, as well as on test-takers' attitudes and acceptance of the use of computers in testing. Furthermore, a study such as this will hopefully assist in the evaluation of computer-based testing, in order to minimize factors which may inhibit optimal achievement performance. If computer-based testing is well accepted by students, and with improvements continually being made in the design of software, it may become a viable method of assessment in the school.

Chapter 2 REVIEW OF LITERATURE

Although the primary uses of microcomputers in education are instructional and administrative, the expansion of computer technology has created many possibilities for computer applications in the area of testing and assessment. McBride (1985) anticipated large-scale applications of computerized testing as computers decreased in cost and became more available. Many important issues have to be considered when administering tests by computers. Among these are the equivalence of scores obtained in computerized testing compared with conventional paper-and-pencil tests, and the impact of computerization on the test-taker. This chapter discusses these issues as well as the current applications of the computer in testing, advantages and disadvantages of computerized testing, and the effects of administering tests via the computer.

2.1 Current applications of the computer in testing

The computer is currently being used in many areas of testing and assessment. In addition to the already established uses of computers for test scoring, calculation of final grades and test score reporting (e.g. Brezezinski, 1984; Turner, 1987), computers can also be used for the determination of test quality (Nelson, 1984), test item banking and test assembly (Hambleton, 1984), as well as for test administration (e.g. Ward, 1984; Millman,1984).

2.1.1 Test and item analysis

Assessing test quality generally involves both item and test analysis. Classical statistics used to summarize item quality are based on difficulty and discrimination indices; these are calculated more easily and quickly with the use of the computer than by traditional hand methods. Items which have been inadvertently mis-keyed, have intrinsic ambiguity, or have structural flaws such as grammatical or contextual clues that make it easy to pick out the correct answer, can be identified and culled out. These items are characterized by being either

too easy or too difficult, and tend to have low or negative discrimination. Test analysis can also provide an overall index of reliability or internal consistency, that is, a measure of how consistently the examinees performed across items or subtests of items.

2.1.2 Item banking and test assembly

Another important use of the computer in testing has been the creation and maintenance of an item pool. This is known as item banking. Hambleton (1984) defines an item bank as "a collection of test items uniquely coded to make the task of retrieving them easier. If the items are not categorized, they are merely a pool or collection of items, not an item bank." In the use of item forms (Hively, Patterson, & Page, 1968; Millman & Outlaw, 1978) which are an alternative to item banks, algorithms are used for randomly generating test items from a well-defined set of item characteristics; each item is similar in structure. For instance, items might have a multiple-choice format, a similar stem, the same number of answer choices, and a common pool of distractors. The most important advantage gained from storing item forms is that many more items can be produced by the microcomputer than would be reasonable to store on the microcomputer (Millman & Outlaw, 1978). With the availability of item forms, unique sets of test items can be developed and drawn for each examinee. Such a feature makes it feasible to administer different tests of the same content and domain to students at different times.

One of the principal advantages of microcomputer-based test development is the ease with which test assembly can be done with the appropriate software. Desirable attributes of an item banking and test assembly system include easily retrievable items with related information, an objective pool, automatic generation of tests, analysis of item performance data, and automatic storage of that data with the associated items (Hambleton, 1984).

2.1.3 Test administration

The computerized administration of tests has also been considered as an attractive alternative to the conventional paper-and-pencil mode of administration. In a computerized test administration, the test-taker is presented with items on a display device such as a cathode-ray tube (CRT) and then indicates his or her answers on a response device such as a standard keyboard. The presentation of test items and the recording of the test-taker's responses are controlled by a computer. Most of the attention to computerized test administration however, has been directed towards psychodiagnostic assessment instruments such as psychological tests and personality inventories. Even in the case of education-related ability and achievement tests, testing (as part of computer-assisted instruction or computer-managed instruction) has mostly been used as the basis for prescribing remedial instructional procedures to determine if the student has achieved mastery, and also to provide the student with some feedback of how he or she performed.

Four main computer-administered testing procedures used in educational assessment settings include computer-based testing, computer adaptive testing, diagnostic testing and the administration of simulations of complex problem situations. Computer-based testing (CBT) generally refers to "using the computer to administer a conventional (i.e. paper-and-pencil) test" (Wise & Plake, 1989). That is, all examinees receive the same set of test items.

Unlike conventional testing where all test-takers receive a common set of items, computer adaptive testing (CAT), or "tailored testing", is designed so that each test-taker receives a different set of items with psychometric characteristics appropriate to his or her estimated level of ability. Aside from the psychological benefits of giving a test that is commensurate with the test-taker's ability, the primary selling point of adaptive testing is that measurements are more precise when examinees respond to questions that are neither too hard

nor too easy for them (Millman, 1984). This test involves making an initial ability estimate and selecting an item from a pool of test items for presentation to the test-taker. According to Green, Bock, Humphreys, Linn, & Reckase (1984), each person's first item on an adaptive test generally has about medium difficulty for the total population. Those who answer correctly get a harder item; those who answer incorrectly get an easier item. After each response, the examinee's ability is re-estimated on the basis of previous performance and a new item is selected at the new estimated ability level. The change in item difficulty from step to step is usually large early in the sequence, but becomes smaller as more is learned about the candidate's ability. The testing process continues until a specified level of reliability or precision is reached and the testing process is terminated. This testing is based on Item Response Theory (Lord, 1980) "which provides the mathematical basis for selecting the appropriate question to give at each point and for producing scores that are comparable between individuals" (Ward, 1984). Adaptive testing allows the tailoring of the choice of questions to match the examinee's ability, bypassing most questions that are inappropriate in difficulty level and that contribute little to the accurate estimation of the test-taker's ability. Individuals with low ability never encounter the difficult questions on which they would resort to blind guessing; others do not waste time with easy questions that they would almost certainly answer correctly. Computer adaptive tests have been shown to take less than half of the time as traditional achievement tests and to provide more precise ability estimates from high to low ability (Weiss, 1983; McKinley & Reckase, 1984, cited in Olsen, Maynes, Slawson, & Ho, 1989).

Another promising use of computer-administered testing is in the area of diagnostic testing. McArthur and Choppin (1984) describe the approach to educational diagnosis as "the use of tests to provide information about specific problems in the performance of a task by an individual student, information that will point to some appropriate remedial treatment" (p. 391). Diagnostic testing is based on the identification and analysis of errors

exhibited by students. Analysis of such misconceptions can provide useful information in evaluating instruction or instructional materials as well as specific prescriptions for planning remediation for a student. Research in this area has mainly been in mathematics education. According to Ronau (1986), "a mistake is an incorrect response, whereas an error is a pattern of mistakes indicating a misunderstanding of a mathematical operation or algorithm" (p. 206). It is believed that a student's systematic errors, which are commonly known as "bugs" are not random but rather are consistent modifications of the correct procedure. The microcomputer has been used to provide a rapid analysis of errors and a specification of the errors that a particular student is making.

The work of Brown and Burton (1978) provides an example of this application. The computer program DEBUGGY can diagnose a number of erroneous rules resulting from misconceptions in whole number subtraction problems. It attempts to identify incorrect algorithms that consistently replicate students' incorrect responses. Included in the program's domain expert is information regarding common arithmetic bugs or procedural errors. The results of all the bugs are compared with the student's answers. A match between the student's answer and the computer generated errors identifies the incorrect strategy the student is using. Similar expert systems such as SIGNBUG have been developed by Tatsuoka, Baillie, & Yamamoto (1982, cited in Tatsuoka, 1983) for diagnosing a number of erroneous rules in signed-number addition and subtraction problems, as well as by Attisha and Yazdani (1984) for use with addition, multiplication and subtraction.

A final current application of computer-administered testing is in the presentation of branching problem simulations. This method however, is not used widely in educational settings but rather in medicine and other health-related fields in professional licensing and certification testing.

2.2 Advantages and disadvantages of computerized testing

The potential benefits of administering conventional tests by computer ranges from opportunities to individualize assessment, to increases in the efficiency and economy with which information can be manipulated. Several of these advantages offered by computerized test administration over printed test administration have been described by Ward (1984), Fletcher & Collins (1986), and Wise & Plake (1989).

Much of educational testing has traditionally been managed on a mass production basis. Logistical considerations have dictated that all examinees be tested at one time. The computer as test administrator offers an opportunity for more flexible scheduling; examinees can take tests individually at virtually any time. During testing, examinees can also be given immediate feedback on the correctness of the response to each question. Computer-based tests, and particularly computer adaptive tests, have been shown to require less administration time than conventional tests (English, Reckase, & Patience, 1977; Olsen *et al*, 1989). For example, using achievement tests with third and sixth graders, Olsen *et al* reported that the computerized adaptive tests required only one-fourth of the testing time required by the paper-and-pencil administered tests, while the computer-based tests required only half to three-quarters of the testing required by the paper-and-pencil administered tests. Hence, when computerized tests are used, students can spend more time engaged in other instructional activites, and less time taking tests.

Another advantage of computerized testing is the capability to present items in new, and potentially more realistic ways (Wise & Plake, 1989). A printed test has display limitations. While it can present text and line drawings with ease, it cannot provide timing of item presentation, variable sequencing of visual displays, animation or motion. The graphics and animation capabilities of computers provide the possibility of presenting more realistically simulated actions and dynamic events in testing situations. Assessment of science process or

problem-solving skills, in particular, are areas where this type of application can be useful (Hale, Oakey, Shaw, & Burns, 1985). Variables can be manipulated and the corresponding outcomes portrayed as they are measured. What results is a more accurate portrayal of situations that rely less heavily than conventional assessment procedures on verbal understanding. For example, the change in length of the shadow cast by a stick at various times of the day can be observed. On a physics test, instead of using a completely worded text or a series of static diagrams to present an item concerning motion, a high-resolution graphic can be used to depict more clearly the motion in question. This should represent a purer measure of the examinee's understanding of the motion concept because it is less confounded with other skills such as reading level. This implies a higher degree of validity for the computerized test item. Computer-animated tests such as this, may have special applications with students who have reading comprehension problems or difficulty translating words into images. Printed tests may therefore not provide an accurate measure of the true ability of the student.

Computer-administered tests offer potentially significant reductions in several classes of measurement error (Bunderson, Inouye, & Olsen, 1989). A mark-sense sheet commonly used in manual testing requires the examinee to code each response by associating it with the item number and then marking one of several bubbles. The visual matching of item numbers and alternative numbers or letters not only takes time but may produce errors. The elimination of answer sheets in computer-administered tests can eliminate some traditional errors such as penciling in the answer to the wrong item number, failing to erase an answer completely, and inadvertently skipping an item in the test booklet but not on the answer sheet. By presenting only one item per screen, the computer automatically matches responses with the item number; examinees can also focus on one item at a time without being distracted, confused, or intimidated by the numerous items per page for paper tests. Computerized tests may therefore provide more accurate measures of performance for students who have lower reading

ability, lower attention span, and higher distractibility. Moreover, convenient features for changing answers can replace time-consuming erasing on printed answer sheets.

The administration of tests by computer also allows the collection of data about examinee response styles. These include information such as which items are skipped, how many answers are changed, and response latencies. The latter may refer to the time it takes an examinee to answer an item; analysis time for any complex drawing, graph, or table; reading time for each option; response selection time, or response speed. Precise measurement of any of these latencies is virtually impossible with paper-and-pencil tests.

Other attractive features of computerized testing include more standardized test administration conditions and immediacy of score reporting. Within a few minutes after completing the test, the examinee or the test administrator can receive a score report and prescriptive profile. Computerized testing also provides for increased test security (Bunderson *et al*, 1989). There are no paper copies of the tests or answer keys to be stolen, copied or otherwise misused. The computer-administered test can include multiple levels of password and security protection, to prevent unauthorized access to the testing materials, item banks or answer keys.

Despite the many advantages associated with computer-administered tests, potential problems exist as well. Use of the response entry device, whether keyboard, touch screen, or mouse can introduce errors. Pressing a wrong key in response to a question results in an error, and the validity of the individual's results is compromised. The amount of printed text that can be shown on a monitor screen can limit both the length of the question and possible responses. The need for multiple computer screens to read lengthy comprehension items might introduce a memory component into the construct being measured (Bunderson *et al*, 1989).

Another problem involves the time lag between an individual's answer to an item and the resulting response from the computer. Long time lags between responses can result in negative user attitudes, anxiety and poor performance (Miller, 1977). Another source of anxiety for individuals using a computer concerns their often mistaken perception that the system will require an inordinate amount of mathematical or computer skills to operate, or that the system can be easily harmed if an error is made by the user (Sampson, 1983). Anxiety and the possible resulting negative impact on performance can occur as a result of poor system design or inaccurate user perceptions or both. A further shortcoming of computer-administered tests, especially in psychodiagnostic assessment, concerns the use of norms in the interpretation of test scores. Most of the tests that are currently administered by computer were originally developed for a traditional paper-and-pencil approach. Differences in mode of administration may make paper-and-pencil norms inappropriate for computer-administered tests.

There are also measurement problems associated with the use of computer-administered tests. These are related to item types, item contamination that arises from certain test design strategies, and the non-equivalence of comparison groups in item analyses (Sarvela & Noonan, 1988). With regard to item type, difficulties arise when constructed-response items (such as fill-ins and short answers) as compared to selected-response items (for example multiple-choice, matching and true/false) are developed for the computer. It becomes almost impossible to program all the possible correct answers, when considering alternative correct answers, wording, spacing and spelling errors. A tremendous amount of programming is involved for even a partial subset of all possible correct answers. There are psychometric implications as well. Students could supply correct answers that simply are not recognized by the computer; the result could be lower reliability and poorer discrimination indices. Because of these reasons, computer-administered tests are mainly restricted to multiple-choice items.

Another psychometric issue in computer-administered testing is the problem of item contamination if instructional design capabilities are incorporated. It is then possible to allow students to preview test items, receive feedback on the correctness of their answers while items are still being presented, or retake items which were drawn randomly from an item pool. In this situation, items which are dependent upon each other (for example, an item which requires the student to use the result from item 3 to compute item 4) would be contaminated if a student receives feedback after each item. Or, the correct answer for one item could provide subtle clues to the correct answer on another item. There are motivational concerns as well. If a student is consistently answering items incorrectly, the negative feedback might be detrimental to motivation on future items. Likewise, a series of correct-answer feedbacks can promote greater motivation in future items. The problem is in the differential effects of item feedback across high and low achieving students. One other contamination problem results from the practice of selecting items randomly from an item bank for a particular test. There is a possibility that a student may see the same items on a second or third try. This problem is exacerbated when item feedback is given. If item feedback is provided, subsequent attempts at tests should contain new items.

Furthermore, when test items are drawn randomly from an item pool, for a given test different students may see different items or items presented in a different order. Consequently, there is non-equivalence of comparison groups. Unless the items administered to one student are equal in difficulty to items that are presented to another student, it becomes extremely difficult to compute item and test statistics (for example, total score, point biserial coefficient, estimate of reliability). The problem is that there is no sensible total score. With random item selection, a total test score is defensible for item analysis only if every item is of equal difficulty and equal discrimination.

2.3 Effects of administering tests via computers

2.3.1 Score equivalence between paper-and-pencil and computer-administered tests

When a conventional paper-and-pencil test is transferred to a computer for administration, the computer-administered version may appear to be an alternate form of the original paper-and-pencil test. However, the scores achieved with computer presentation may not necessarily be comparable to those obtained with the conventional format, and empirical verification is necessary before a claim of equivalent validity is justified. Even though the content of the items is the same, mode of presentation could make a difference in test-related behaviors, such as the propensity to guess, the facility with which earlier items can be reconsidered, and the ease and speed of responding (Greaud & Green, 1986). Duthie (1984, cited in Wilson, Genco, & Yager, 1985) has suggested that there may be cognitive differences in the manner in which a person approaches computer-administered and paper-and-pencil testing tasks. The manipulation necessary for working with a computer, and the stimulus value of the computer itself may alter the manner of cognitive functioning exhibited by the test-taker. Wood (1984) and Duthie have both noted that test performance may well be influenced by such seemingly minor differences as the formatting of a microcomputer screen display.

The concern for score equivalence is acknowledged in the American Psychological Association's (APA) *Guidelines for Computer-based tests and Interpretations* (1986). Guideline 16 states that:

> When interpreting scores from the computerized versions of conventional test, the equivalence of scores from computerized versions should be established and documented before using norms or cutting scores obtained from conventional tests. Scores from conventional and computer administrations may be considered equivalent when (a) the rank orders of scores of individuals tested in alternative modes closely approximate each other, and (b) the means, dispersions, and shapes of the score distributions are approximately the same, or have been made approximately the same by rescaling the score from the computer mode (p. 14).

One way to look at the issue of empirical validation of an equivalent form of a test is from the point of parallel tests in classical test theory. Following from the definition of parallel tests, the subtest and total test scores for a paper-and-pencil test and its computer administered counterpart should yield equal means, equal variances, and equal correlations with the scores on any other criterion variable (Allen & Yen, 1979; Ghiselli, Campbell, & Zedeck, 1981). If the scores from the computer-administered test version are intended to be interchangeable with scores obtained by the paper-and-pencil test, then the two test versions can be evaluated against the criteria for parallel tests.

Green *et al* (1984) have suggested some possible ways in which the psychometric characteristics of tests might be altered when items are switched from paper-and-pencil to computer administration. First, there may be an overall mean shift resulting from a change in the difficulty of the test, with the items being easier or harder. Tests of speed performance in particular, where response time is a determining factor, would be expected to show an overall mean shift, because the time to respond depends critically on the nature of the response. Second, there could be an item-by-mode interaction. Some items might change, others might not, or some might become harder, others easier. This would be most likely to occur on tests with diagrams; the clarity of the diagrams might be different on the screen. Items with many lines of text, such as paragraph comprehension items, might also show this effect. Third, the nature of the test-taking task might change. For example, students who are more familiar with computers may perform somewhat better on the computer-administered version of the test than equally able students who are less familiar with computers. As a result, the test may unintentionally measure computer literacy along with the subject matter.

Several factors influencing the equivalence and psychometric properties of tests from the two formats have been proposed. One variable that has been used to explain medium effects or differences on examinee scores is the differences in test-taking flexibility and amount

of control (Spray, Ackerman, Reckase & Carlson, 1989). This refers to whether examinees are allowed to skip items and answer them later in the test, return to and review items already answered, and change answers to items. If computerized versions of tests do not provide these features and instead display individual items in a single-pass, no-return mode, then this may result in differences in item characteristics, such as the difficulty and discrimination indices. Individual differences in test anxiety, computer anxiety and attitudes toward computerized testing, and amount of previous computer experience have also been hypothesized to affect the comparability of scores (Llabre *et al*, 1987). If these variables differentially affect examinee performance to a significant degree, then they may have implications for equity issues in testing. Other factors that have been suggested to affect the equivalence of scores include the difficulty of the test and the cognitive processes required by the test (Lee, Moreno, & Sympson, 1986), as well as test structure (discrete items versus sets of items based on a common reading passage or problem description), item content (items containing graphics versus items containing only verbal material), test timing (speed versus untimed tests), and item feedback on the test performance (Mazzeo & Harvey, 1988).

2.3.2 Studies of the effect of medium of item presentation on test performance

Studies that have investigated differences in examinee performance on items administered in paper-and-pencil format or via a computer have produced equivocal results. A recent review of the existing research on this issue of score equivalence between conventional and computer-based tests, commissioned by the College Board and Educational Testing Service was conducted by Mazzeo & Harvey (1988). Although there are numerous studies which indicated equivalence of scores and reliabilities, there are also many other studies showing significant mean differences between the scores from the two modes of testing. In general, it was found more frequently that the mean scores were not equivalent than that they were equivalent; that is, the mean scores for tests administered on paper were more often higher than

for computer-administered tests. However, the score differences were generally quite small and of little practical significance.

An obvious difference between paper-and-pencil and computerized presentation modes is the method of responding. On paper-and-pencil tests, comprising multiple-choice questions for example, examinees usually respond by locating and filling in a bubble on an answer sheet; on computer-presented tests, examinees press a response key on a keyboard. Pressing a computer terminal key rather than marking a bubble on an answer sheet may affect response time. If response time is a critical component of an examinee's score on a test, then this difference in method of responding is likely to affect test performance. This issue of response speed is particularly important on speed clerical tests. A speed test consists of items that all examinees could answer correctly given enough time, but the test is given with a short time limit to see how quickly examinees can work (Allen & Yen, 1979). One might expect, *a priori*, that speed tests would be particularly sensitive to mode-of-administration effects. Such effects could be facilitative, if the manual operations associated with recording answers were less cumbersome in computer-administered versions of the test. On the other hand, negative mode-of-administration effects may be observed if the computer version is more awkward to work on. However, for most cognitive power tests of aptitude or achievement, the time to indicate an answer may be an inconsequential factor because most of the time is spent deciding which answer to choose. In a power test, each examinee, depending on his or her ability can correctly answer only a certain number of items even without time limits. Often, time limits are generous to ensure that each examinee can attempt each item (Allen & Yen, 1979).

In a study by Lee *et al* (1986), the performance of 585 military recruits on a computerized version of an Experimental Arithmetic Reasoning test (EXP-AR) was compared with that on a pencil-and-paper version of the same test. The EXP-AR was developed from items similar to those used in the Armed Services Vocational Aptitude Battery arithmetic

reasoning test (ASVAB-AR). Each subject was randomly assigned to either the paper-and-pencil administration or the computer administration. No time limit was imposed. Subjects could not refer to previous items or change answers; they were also not permitted to omit items. This flexibility was however, present in the paper-and-pencil group. A significant main effect for mode of test administration was found ($p<0.05$), with the mean score obtained by computer group lower than that obtained by the paper-and-pencil group. The results also showed that item difficulty was affected by mode and the effect was fairly uniform across items. Twenty-one of the 30 items were more difficult in the computer mode, while 3 were more difficult in the paper-and-pencil mode. A possible explanation for the significantly lower scores in the computer group may be that examinees did not have an opportunity to review previous responses and to correct erroneous responses. Moreover, the ability to view multiple items in the paper-and-pencil mode may have aided performance in that mode.

Eaves & Smith (1986) also compared the effects of item presentation medium on test scores of 96 college students enrolled in an undergraduate educational media class. Subjects were blocked according to their amount of previous experience with microcomputer and then randomly assigned to one of the two testing modes, viz. computerized and paper-and-pencil. Students taking the paper-and-pencil tests were able to move back and forth within the test, scan the test as a whole, review items and change answers, but those taking the computer-based test were required to deal with only one stimulus item at a time, and once they responded to an item, they were unable to change the response. In addition, they could not scan the entire test, skip items, nor omit an item accidentally. Although differences existed in the response demands and test-taking flexibility placed on examinees in the two groups and the investigators hypothesized that the results should favour the paper-and-pencil group, no differences between the two item presentation media were found ($p>0.05$), regardless of the level of previous microcomputer experience.

In contrast to the above two studies, Spray *et al* (1989) designed a study where item administration procedures were made as identical as possible for each presentation medium. The computerized tests for military technicians of a Ground Radio Repair Course was written to mimic as closely as possible all of the flexibility of the paper-and-pencil format. Examinees were able to move back and forth through the test, review previous responses, and change answers. The items appeared on the screen exactly as they appeared in the printed version. No figures or graphics were required to be displayed on the screen. Test results on three units of instruction, called "annexes", were analyzed. No significant difference in mean test scores were found for two of the three annexes. For the third annex, the mean score on the paper-and-pencil test was significantly higher ($p<0.05$)

Other studies have also shown no difference in mean test scores as a function of mode of test administration (Olsen *et al* ,1986; Wise & Wise, 1987), and although an equivalency in item administration procedures was not explicitly stated, such a condition might have been partially responsible for the score equivalence as well. As part of a larger study that included computer-adaptive tests, Olsen *et al* compared computer-administered versions of the California Assessment Program mathematics applications test with paper-and-pencil versions for both third-grade and sixth-grade students from several California school districts. In the computer-administered form, the entire set of items in the paper-and-pencil version was presented sequentially on the CRT of a PLATO / WICAT system 300 microcomputer system. Students entered their responses on the computer keyboard. No details were provided in the report regarding whether they could review items or change their answers. No significant differences were found between score means obtained on the paper-and-pencil and computer versions for either the third- or the sixth-grade groups. In addition, the authors computed coefficient alpha reliabilities and average standard errors of measurements for the two versions at each grade level. The results of the paper-and-pencil and the computer administrations were almost identical at both grade levels, differing only in the second decimal place.

Wise & Wise (1987) compared the performance of 68 third-grade and fourth-grade students on paper-and-pencil administered and two computer-administered versions of a 32-item multiple choice test of basic arithmetic skills. In one computer version, immediate feedback on the correctness of a response was provided. In the other computer version, no feedback was provided. Subjects were randomly assigned to one of the three administration modes. Number-right arithmetic skills scores were analyzed in a three modes-of-administration by two levels-of achievement ANOVA. No significant main effects or interaction were obtained.

The results of the aforementioned studies seem to indicate that if score equivalence between item presentation media is required, then such equivalence can be achieved if test-taking flexibility under both conditions is equivalent. But, if the test-taking process is not the same for the two administration procedures, then the test items may function differently. Where scores may be due at least partly to the constraint imposed in the computerized mode on reviewing and changing answers, it is important to provide item administration procedures that are as identical as possible for each presentation medium.

In studies examining the effects of computer presentation of speed tests, results suggest that differences in the method of indicating responses and the number of items presented at a time across the two modes will likely affect the rate of responding, and thus affect the comparability of computerized and conventional test scores. Greaud and Green (1986) compared performance on paper-and-pencil and computerized versions of two clerical speed tests from the ASVAB (Armed Services Vocational Aptitude Battery), the numerical operations test (NO) and clerical speed test (CS). The NO test consisted of four-option multiple choice arithmetic problems. The CS test consisted of a single master list of words, with each word adjacent to a four-digit numeric code. Below the master list was a second list of five-option multiple-choice items. Each item contained one of the words in the master list and five numeric

codes. An examinee's task was to select the code that corresponded to the word in the master list. For the paper-and-pencil versions, problems were presented on a single sheet of paper. For the computer version, each item was presented individually on the computer.

A significant effect for mode of test administration on scores was observed for both tests. Scores increased from paper-and-pencil administration to computerized administration by 37% for the NO tests and by 62% for the CS test. Examinees were very much faster at taking these speeded tests on the computer than with paper-and-pencil. Reliability coefficients for the NO test were significantly higher ($p=0.05$) in the computerized version (0.77) than they were in the paper-and-pencil version (0.62). No significant differences in reliability coefficients were observed for the two versions of the CS tests (0.70 for paper-and-pencil versus 0.74 for computer). Furthermore, for the CS test, the between-mode correlations for the two sets of scores were low, ranging from 0.28 to 0.61; this indicated that the task was different in the two modes. On the computer, each item was presented individually, whereas the items were printed on the same page in the paper-and-pencil test. Apparently, this change had an effect on test performance. The implication of this task difference is that care must be exercised in transferring a test from paper-and-pencil to the computer. Performance on tests such as CS can be affected if modifications are made in the way items are presented, that is, individually rather than in a group.

Whether a computer-based test results in scores equivalent to a paper-and-pencil test may also in part, depend on the requisite ability measured by the test. In a study by Sachar and Fletcher (1978, cited in Mazzeo and Harvey, 1988), two types of tests that were administered by paper-and-pencil and by computer were compared. The vocabulary test consisted of 50 five-option multiple choice items and required the examinee to answer an average of 3.33 questions per minute. The symbolic reasoning test which required the examinee to answer an average of 6.00 questions per minute was a 30 item true / false test. The

computerized items were shown one at a time; examinees could skip items, review previous items and change answers. To do so, however, they were required to go back (or forward) sequentially through each previous item to review or change answers to items not adjacent to the one they were currently working on. The overall difference in means in the computer and paper-and-pencil versions was not statistically significant for the vocabulary test but was significant for the symbolic reasoning test. The results of this experiment showed clear mode-of-administration effects for the symbolic reasoning test but not for the vocabulary test. One distinct difference between the tests is the difference in speed involved. The demands of taking the test by computer, in particular the operations involved in reviewing previous items and changing answers, and certain delay inherent in retracing the test to earlier items may sufficiently slow performance, particularly in the highly speeded symbolic reasoning test, to result in decreased scores on the computer version. In sum, the studies described showed mode-of-administration effects for highly speeded tests where accuracy and speed are important, as any distractions or difficulties introduced by computer administration, or advantages introduced by ease of responding, could affect the rate at which items are answered.

In tests that contain graphics (i.e. figural or pictorial information), the principal concern is whether the resolution of the display medium used in a computer-administered version is sufficient to allow the examinee to correctly interpret, and extract the required information from the pictures or diagrams. For timed tests or for speed tests, there is an additional concern whether the rate at which the material that can be assimilated is decreased or increased, resulting in concomitant effects on score comparability. This factor which is not inherent in test items consisting of only numeric and alphanumeric characters, may make pictures more difficult to recognize or interpret and consequently, may have an impact on performance.

Five studies that used tests containing graphics were described by Mazzeo and Harvey (1988). The evidence however, is mixed concerning whether the display of graphics has an effect on the relative difficulty of computer and paper-and-pencil versions of a test. Three of the studies found no evidence for significant differences caused by mode of administration (Reckase, Carlson, & Ackerman, 1986; Kiely, Zara, & Weiss, 1986; Jacobs, Byrd, & High, 1985) while the other two studies did (Wilgrube, 1982; Jonassen, 1986). The results of Wilgrube's study however, are difficult to interpret in that order effects may be confounded with mode-of-administration effects, because all computer tests were administered subsequent to the paper-and-pencil test. In Jonassen's study, different half-tests were administered in each of the two administration modes; hence these differences in difficulty may be confounded with mode-of-administration effects.

A final point worthy of note concerning computer-based tests is that it is often impossible to simultaneously display an item along with a long reading passage or a set of figures to which the item refers. Examinees must route themselves through a number of related screens to obtain information necessary to complete an item; this may be distracting or may make a given task more cognitively demanding. The result could be that computer versions appear more difficult than their paper-and-pencil counterparts. Furthermore, particularly for timed tests, the number of keystrokes involved cause computer versions to be more speeded up and therefore appear more difficult. Being able to scroll through the reading passage while an item is concurrently displayed is a desirable feature to be incorporated in a computer test that contains reading passges.

In summary, on the basis of the above review, a number of factors related to the implementation of computer-based tests have been identified. First, if test-taking flexibility under both item administration media is comparable, then equivalent scores can be achieved. Second, speed tests seem to be particularly sensitive to mode-of-administration effects and

scores from computerized versions will most likely not be comparable with paper-and-pencil versions. The presentation of figural and pictorial information may also have an effect on score equivalence. However, results from studies of tests containing graphics have been inconclusive.

2.3.3 Test Anxiety Studies

Although the primary determinant of examinee responses to items on cognitive tests is knowledge or aptitude, other factors such as test anxiety have been shown to be related to test performance. General summaries of research (Sarason, 1980; Tryon, 1980) and classical specific studies (Mandler and Sarason, 1952) point to moderately negative relationships between test anxiety and test performance.

Dusek (1980) defines test anxiety as "an unpleasant feeling or emotional state that has physiological and behavioral concomitants, and that is experienced in formal testing or other evaluative situations" (p. 88). Test anxiety is a special case of general anxiety and has been conceptualized as a situation-specific anxiety trait. Two meanings of the term anxiety can be distinguished: anxiety as a state and anxiety as a trait. The state-trait model of anxiety set forth by Spielberger (1972a) describes state and trait anxiety as follows:

> State anxiety (A-State) may be conceptualized as a transitory emotional state or conditions of the human organism that varies in intensity and fluctuates over time. This condition is characterized by subjective, consciously perceived feelings of tension and apprehension, and activation of the autonomic nervous system. Level of A-State should be high in circumstances that are perceived by an individual to be threatening, irrespective of the objective danger; A-State should be low in nonstressful situations, or in circumstances in which an existing danger is not perceived as threatening.
>
> Trait anxiety (A-Trait) refers to relatively stable individual differences in anxiety proneness, that is, to differences in the disposition to perceive a wide range of stimulus situations as dangerous or threatening, and in the tendency to respond to such threats with A-State reactions (p.39).

Although test situations are stressful and evoke state anxiety (A-State) reactions in most students, the magnitude of the A-State response will depend on the student's perception of a particular test as personally threatening. Individuals with high test anxiety generally perceive tests as more threatening than low test-anxious individuals and respond with greater elevations in state anxiety to the evaluative threat that is inherent in most test situations.

Correlational studies have shown that the performance of highly test-anxious persons on complex tasks is deleteriously affected by evaluational stressors. Individuals having high scores on measures of test anxiety tend to perform relatively poorly on ability and achievement tests, when compared with low anxiety scorers (Sarason,1972). The generally accepted current explanation of the negative effects of test anxiety is that they result from ineffective cognitive strategies and attentional deficits that cause poor task performance in evaluative situations. Children with low anxiety level appear to become deeply involved in evaluative tasks but highly anxious children do not. Highly anxious children seem to experience attentional blocks, extreme concern with autonomic and emotional self-cues, and cognitive deficits such as misinterpretation of information. The highly anxious child's attentional and cognitive deficits are likely to interfere with both learning and responding in evaluative situations and result in lowered performance. Wine (1971) suggested an "attentional" interpretation of the debilitating effects of test anxiety. She contends that, during examinations, highly test-anxious individuals divide their attention between task requirements and task-irrelevant cognitive activities, such as worry. These worry cognitions distract students from task requirements and interfere with the effective use of their time, thereby contributing to performance decrements. According to Wine, the highly test-anxious person responds to evaluative testing conditions with ruminative, self-evaluative worry, and thus, cannot direct adequate attention to task-relevant variables.

Sex differences in test anxiety have also been consistently obtained, with females having higher levels of anxiety (Spielberger, 1972b, p. 48; Crocker, Schmitt, & Tang, 1988). Given the fact that research has provided evidence of a negative relationship between test anxiety and test performance, an important issue related to the use of computers in testing is whether computer-administered testing will increase test anxiety and depress test performance, particularly in examinees who are relatively unfamiliar with computers. The relationship between anxiety and computer-administered testing was explored by Hedl, O'Neil, & Hansen (1973) when computers were less commonly used in educational settings. College students who took an individually administered computerized intelligence test had higher levels of state anxiety, as measured by Spielberger, Gorsuch, & Lushene's (1970) State-trait Anxiety Inventory both before and after the test, than students given the same test administered by paper-and-pencil. The authors speculated that anxiety may have been produced by procedural variables in the computer-administered test such as unfamiliarity with the terminal operations and the type of interaction that the students were required to have with the computer. The results of this study also showed no evidence of interactions between sex and state anxiety.

More recent investigations to determine the effect of a computer-administered test on test anxiety and performance have also been carried out. In a study by Llabre *et al* (1987), forty college students of whom 77.5% reported never or seldom having used a computer, were randomly assigned to either a computerized or paper-and-pencil version of the California Short-Form Test of Mental Maturity (CMM). Both groups were also given a revised version of Sarason's (1978) Test Anxiety Scale (TAS-R) before taking the test. The results indicated significant differences in anxiety level ($p<0.05$), with the group taking the computerized test displaying higher levels of test anxiety. Significant differences were also obtained on the CMM ($p<0.01$), with the group taking the computerized test group obtaining slightly lower test scores. There was, however, no systematic relationship (linear or

curvilinear) between anxiety and performance. The results of this study should be viewed with caution however, because the sample size for each group was small.

In another study by Ward *et al* (1989), fifty college students were randomly assigned to take an Education class exam either on computer or in the traditional paper-and-pencil manner. Following testing, examinees were administered a questionnaire designed to measure their test anxiety and attitudes towards computerized testing. Results indicated no differences in test performance ($p>0.35$) but a significant difference in anxiety level ($p<0.025$) with those tested by computer having a higher anxiety level. The authors hypothesized that this increase in anxiety might be attributable to the novelty of the computer testing situation or the result of a fear of computers. The results also indicated a negative attitude towards computer testing with 75 % of the computer tested group agreeing that computer testing was more difficult than traditional methods.

Given the results reported in the preceding section, it appears that the added test anxiety associated with computer-administered tests is an important consideration in the evaluation of computerized testing. There is a need to familiarize examinees with the technology used in testing prior to test administration so that anxiety about computers does not increase examinee's level of test anxiety.

2.3.4 Computer Anxiety studies

As noted previously, individual differences in computer anxiety has been hypothesized as a factor affecting the performance of an examinee on a computer-based test. This hypothesis rests on the assumption that examinees must feel comfortable with the computer and confident about their ability to work with a computer before being able to use the computer effectively to take a test. As anxiety towards using computers may influence the

testing process, such an affective reaction may therefore be an important factor in whether computer-based testing becomes an accepted component of the evaluation of a school system.

Computer anxiety is generally perceived as a situational manifestation of a general anxiety construct, fitting into the category of anxiety state rather than anxiety trait. Raub (1981, cited in Cambre and Cook, 1985) defined computer anxiety as "the complex emotional reactions that are evoked in individuals who interpret computers as personally threatening." Simonson, Maurer, Montag-Torardi, & Whitaker (1987) described it as "the fear or apprehension felt by individuals when they used computers, or when they considered the possibility of computer utilization." Manifestations of computer anxiety may thus be triggered by consideration of the implications of utilizing computer technology by planning to interact with a computer, or by actually interacting with a computer.

Factors such as gender and prior computer experience have been identified as being related to computer anxiety. A review of previous research reveals several studies designed to determine sex-related differences in computer anxiety and attitudes. While Loyd and Gressard (1984a) found no difference in computer anxiety levels for males and females in a sample of high school and college students, Chen (1986) on the other hand, found significant sex-related differences, with high school males being less anxious and holding more positive attitudes of interest in and confidence with computers than did females. Differences in computer attitudes such as interest, liking and confidence were also obtained in investigations by Levin and Gordan (1989), and Popovich *et al* (1987) with males holding more positive attitudes.

The amount of experience with computers is also a significant factor in computer anxiety because anxiety is produced in part, by a lack of familiarity with computer use. In fact, a major finding of the study by Levin and Gordan (1989) suggested that prior computer

exposure has a stronger influence on attitudes than does gender. Students with little or no computer experience were significantly more anxious about computers than those with more experience. This finding is supported by Loyd and Gressard (1984a) who found that although students' attitudes towards computers were not dependent on sex, they were affected by the amount of computer experience, with more experience related to decreased anxiety and increased positive attitudes. Manifestations of computer experience could be having access to a computer at home, participating in computer-related courses, playing computer games or knowing how to work with computers. Students who have a computer at home tend to have lower computer anxiety than those who do not (Hayek and Stephens, 1989; Johanson, 1985). Boys are also more likely to have used computers more frequently at both home and school, as well as in informal settings (Chen, 1986). Perhaps because of this, they are often found to be less anxious about using computers and more self-confident about their abilities with computers.

A study which examined the effect of computer anxiety on test performance was conducted on a sample of college students by Wise *et al* (1989). Since computer anxiety might negatively affect one's performance, this variable was hypothesized to exacerbate score differences between computer-administered and paper-and-pencil testing modes. Contrary to expectations, the authors found that computer-anxious examinees did not have significant score differences between computer-based and conventional tests. A possible explanation of this unexpected finding may be that if the demands made on the examinee by the computerized mode of testing are not too complex and the tasks kept simple, any computer anxiety felt by the examinee may not lower test performance significantly. Earlier research by Denny (1966), however, showed anxiety to be related to poorer performance on a computerized test. As the results of the studies on the relationship between computer anxiety and test performance are mixed and inconclusive, further research in this area is warranted.

2.3.5 Effects of computer experience on computerized test performance

Another individual difference variable, the amount of previous computer experience has also been hypothesized to have an effect on computerized test performance. Inexperience and unfamiliarity with computers may increase anxiety and interfere with test-taking. If this were the case, then computerized testing may discriminate against examinees who have not worked with computers prior to testing. Those who have more past experience with computers would then be at an advantage when taking a computerized test. Thus individual differences in terms of past access to computers may be an important issue in computer-based testing.

Previous research has shown that the amount of computer experience can influence test performance on computer-based tests (Johnson & White, 1980; Lee, 1986), with less experience being associated with lower test scores. Johnson and White used a between subjects design to compare computerized test scores of a sample of elderly subjects who had prior training on the computer with the scores of those who did not have prior training. They found that increased training on the computer prior to testing significantly enhanced the test scores of their examinees. The authors attributed the improvement in scores to the amelioration of anxiety by the training. Lee's study investigated the performance on a computerized arithmetic reasoning test with a sample of college undergraduates. While past computer experience was a significant factor affecting test performance, the findings showed that there was not a significant difference between "low experience" and "high experience" persons, indicating that minimal work with computers may be sufficient to prepare a person for computerized testing. Furthermore, it was also found that those whose computer experience involved computerized games only, performed significantly worse than the other two groups, indicating that computerized games did not provide the same training with computers as work tasks.

Contrary to the above findings, the results of three other studies showed that lack of experience with using computers did not have an adverse effect on examinee performance on a computer-based test. The subjects in the sample pool in these three separate studies by Eaves & Smith (1986), Plumly & Ray (1989), and Wise *et al* (1989) were all college students. There are some plausible reasons why contradictory findings were obtained. First, age may play a part in the ability of examinees to respond equally to the two media used in the studies, namely computerized and traditional paper-and-pencil tests. It would seem reasonable to assume that college students would be more likely than elderly examinees to adapt to the novelty of using computers in testing. Second, the response demands placed on the subjects for the tests in the latter three studies might have been simple enough that an examinee with little or no prior computer experience would not be disadvantaged by the computerized test-taking procedures.

2.3.6 Examinees' reactions to computerized testing

To date, there has been little research regarding students' reactions to computerized testing. The research literature on attitudes toward computerized assessment has primarily focused on the reactions of examinees in the clinical and psychodiagnostic realm. However, a few researchers have investigated the reactions of examinees toward aptitude and achievement tests. In these studies the reactions of the test-takers were generally favourable.

In the study by Gwinn & Beal (1988), 70 % of the university students who took an anatomy and physiology test had a decided preference for computer testing over paper-and-pencil tests, about 7 % disliked it, and the remainder found it made little difference. This sample of students had very little prior experience with the use of computers. A greater preference for on-line computer testing was also found by Moe and Johnson (1988) who investigated the reactions of Grade 8 to 12 students on a standardized aptitude test battery. Overall reactions to the computerized test were overwhelmingly positive; 91 % of the subjects indicated they would choose a computerized test. Nearly half of the students reported that they

experienced no problems during the computerized test. Of those who did report trouble, the major difficulty was with the computer screen; 63 % said their eyes got tired, 39 % indicated that the screen was too bright, and 27.6 % were disturbed by the glare on the screen. Most students (88.5 %) however, said they had no difficulty with using the keys. When asked for the "worst things" about the test, the two most serious complaints were glare and the lack of opportunity to review answers. The most common response for the "best things" about the computerized test was the ease of answering. Other popular responses were that the test seemed faster, and that the computerized test was "fun".

Fletcher and Collins (1986) conducted a survey on university students taking a Biology test for their relative preferences for the two forms of the tests. The most often cited criticisms about computerized tests were the inability to skip questions and answer them later, and the inability to review answers at the end of the test and make changes. Furthermore, with such constraints, examinees could not get hints for responses from other questions. Despite these criticisms, most of the respondents preferred the computer-administered test, and cited the immediacy of scoring, increased speed of test-taking and immediate feedback on incorrect answers as the major advantages.

While the above studies reported generally positive attitudes toward computerized tests, the college examinees in the study by Ward *et al* (1989) exhibited a negative attitude toward computer-based testing. Seventy-five percent of the computer tested group agreed that computer testing was more difficult than traditional methods.

2.4 Summary

The results of studies comparing the effects of computer-based tests with traditional paper-and-pencil tests suggest that differences in performance between the two modes of testing might depend upon the type of test taken and the population tested. A number of

variables, both procedural and examinee-specific can moderate the effects of computerized test-taking on examinee performance and affect the score equivalence of certain individuals or groups of individuals. These factors include the response demands required in the test-taking process, test anxiety, computer anxiety, and the amount of previous computer experience. Issues such as the comparability in test-taking flexibility and examinee control are important considerations in the implementation of computer-based tests. The effect of prior computer-related experience on test performance may be reduced by making the response demands to the computerized tasks simple.

Chapter 3 METHODOLOGY

3.1 Overview

The purpose of this study was to compare the relative performance and test anxiety level of Grade 10 students taking a conventional paper-and-pencil science achievement test with those taking a computer-based version of the same test. A 2 X 2 (mode of test administration by gender) factorial design was used as the experimental design for the study. The study also assessed if there was any interaction between mode of test administration and gender, and whether the amount of previous experience with computers affected the computer anxiety level of students tested by the computer. In what follows, the selection of subjects, data-gathering instruments, research design, experimental procedure and data collection, techniques used for data analysis, and methodological assumptions are described.

3.2 Selection of subjects

The subjects for the study were selected from a secondary school in the Burnaby, B.C. school district. This school was selected because of the accessibility to Macintosh microcomputers; this was a necessary prerequisite for the implementation of the study. A total of 105 students (54 males and 51 females) from five of the school's six Grade 10 science classes participated in the study. The sixth class was excluded because of time-table scheduling constraints. As ability grouping was not practised in the school, the science ability range of the students from the five participating classes could be considered equivalent.

3.3 Instrumentation

3.3.1 Science achievement test (paper-and-pencil version)

A thirty-four item science achievement test was developed, utilizing the 1986 British Columbia Provincial Science Assessment Grade 10 item bank. This test is found in Appendix A. The item bank consisted of 120 questions selected from items previously used or

piloted in the 1978 and 1982 Science assessments; items from other sources including previously used B.C. province-wide tests, Manitoba Assessments, the Ontario Assessment Instrument Pool, California Assessments, National Assessment of Education Progress (NAEP), the 1984 International Science Study of IEA, several different standardized tests; and original items generated for the 1986 assessment (Bateson, Anderson, Dale, McConnell, & Rutherford, 1986). The items were designed to measure achievement on the total Junior Secondary (Grades 8-10) curriculum.

The items on the science test were classified under four domains and three topics. These domains and topics were established by the 1986 Science Assessment Committee. They reflected the content and learning outcomes described in the current Junior Secondary curriculum as well as part of the curricula used in 1978 and 1982. The four domains are "Processes and skills", "Knowledge: recall and understand", "Application of science concepts" and "Rational and critical thinking". The three topics are "Physical sciences", "Life sciences" and "Earth / Space sciences". "Processes and skills" involved observing, classifying and interpreting information for the purpose of solving problems, demonstration of laboratory skills, interpreting data and making predictions based on information given. An example of such a question would be test item number 30, "As the magnification of the microscope increases, what happens to the width of the field of the image?". "Knowledge: recall and understand" involved the ability to recall and understand various science facts, concepts, and principles. Test item number 29, "What does the Milky Way consist of?" is an example. "Application of science concepts" referred to applying relevant scientific knowledge and methods to a new problem. Test item number 2, "In guinea pigs, fur colour is dependent on only one pair of genes and black is dominant over white. If no mutations occur, what will happen if a purebred black guinea pig is crossed with a purebred white guinea pig?" falls into this category. "Rational and critical thinking " referred to the ability to solve problems by transferring prior knowledge and /or learning behaviour. This included integrating learning from different areas to solve a

problem, developing alternate solutions to a given problem, reasoning abstractly, and critically evaluating scientific issues. An example is test item 20, "How can excessive exposure to radiation affect future generations?".

Based on these domains and topics, a table of specifications was constructed. Proportional sampling was employed to select items from the item pool. This maintained the balance of items by domain and topic when compared to the original item pool. Only items with a difficulty index or p-value in the range of 0.31 to 0.69 were considered. This eliminated items that were very easy or very difficult. The aim was to construct a test which had an average item difficulty of 0.50 as this would give maximum variance among scores and hence maximum discrimination among test-takers (Crocker and Algina, 1986, p.98; Ghiselli *et al,* 1981, p. 430). All items which met this range criterion were then arranged in descending order of difficulty. Systematic random sampling was used to select the questions for the test. That is, every third question in the list was chosen. It was also ensured that the content of the questions was balanced. For example, if a question chosen by systematic random sampling had similar content to a previous question, the next question in the list would be selected instead. The table of specifications illustrating the structure of the test and the p-values is shown in Table 1. The numbers at the bottom left hand corner of each cell represent the mean cell p-value; the numbers of items for each cell are indicated at the right hand corner. The overall mean p-value of the thirty-four item test was 0.50. A second table of specifications describing the organization of the test items is included in Appendix B.

All items in the test were multiple-choice questions with five alternatives. The questions were presented in descending order of difficulty with two questions on each printed page. To ensure that the presentation format of the questions was as similar as possible to the computerized test, the paper-and-pencil test booklet was a printed copy of the computer-based test. Hence, the appearance of each question on the printed page and the computer screen was

Table 1 **Table of specifications showing p-values and number of items for science achievement test**

Domain	Topic						Total	
	Physical sciences		Life sciences		Earth/space sciences			
Skills & Processes	0.54	5	0.48	2	0.49	1	0.52	8
Knowledge -Recall & Understand	0.47	4	0.54	4	0.57	2	0.52	10
Application of science concepts	0.48	4	0.47	4	0.58	2	0.50	10
Rational & critical thinking	0.46	2	0.50	3	0.40	1	0.47	6
Total	0.49	15	0.50	13	0.53	6	0.50	34

Note. Numbers at the bottom left and right hand corners of each cell represent the mean p-values and number of questions respectively.

The overall mean p-value of the test is 0.50.

identical. Examinees had to write their responses (a, b, c, d, or e) into a box at the bottom of each question frame. The time limit for the test was 40 minutes. As the science achievement test was designed to be more of a power test than a speed test, this time limit was generous enough to ensure that each examinee could attempt every item.

3.3.2 Science achievement test (computerized version)

The computer-based test was developed to be identical in length, item content and sequence to the paper-and-pencil test. It was programmed in Hypercard, version 1.2.2 (Apple Computer, 1988), and administered via the Apple Macintosh microcomputer. The same thirty-four items were presented one at a time, and each test question was displayed on the computer screen exactly as it appeared on the printed form. Examinees indicated their responses by typing either a, b, c, d, or e from the keyboard into a box at the bottom of each question frame and pressing the *Return* key. In designing the computer-based test, every attempt was made to ensure that the test-taking flexibility was as equivalent as possible to the conventional paper-and-pencil test. Examinees could move back and forth within the test by pressing the *left* and *right* arrow keys respectively, items could be skipped as well as answered in any order, and responses could be reviewed and changed any number of times.

Although the item content and procedural features were similar to the paper-and-pencil test, there was a possibility that a difference existed in the response demands placed on the examinees. In order to return to a previous question or move forward to a later question, examinees had to move through consecutive items to reach that particular question. Another difference was that the questions were presented singly. When the examinees had decided that they were finished with the test, they terminated the testing session by typing "Q". The number of keystrokes involved was kept to a minimum. Examinees were not required to use the Macintosh mouse at all as those unfamiliar with its operation might be disadvantaged when responding to the test items. The program recorded the names of the examinees and their

responses, scored the responses, and stored all collected data for subsequent item and test analyses. The time limit for the computerized test was also 40 minutes.

3.3.3 Attitude questionnaire

The Attitude questionnaire was used as a premeasure of an examinee's level of test anxiety and computer anxiety. It consisted of two sections. Section A , "Self-evaluation questionnaire" was Form X-2 of the State-trait Anxiety Inventory (STAI) developed by Spielberger, Gorsuch, & Lushene (1970). Section B, "Attitudes towards computers" was the Computer Anxiety subscale of the Computer Attitude Scale (CAS) developed by Loyd and Gressard (1984b).

Form X-2 of the STAI which was the A-trait scale consisted of 20 items designed to assess trait anxiety. Items were presented in counterbalanced order relative to anxiety; that is, there was an interspersing of positive and negative statements. This was done to minimize the possibility of a response set. Subjects responded to the statements by selecting one of four responses ranging from "Almost never" to "Almost always". Scores could range from 20 to 80. The scoring keys reversed the direction of the nonanxiety items so that a high score suggested high trait anxiety. Test-retest reliabilities for the scale for males and females over a 104-day period were 0.73 and 0.77 respectively (Dreger, 1985). Alpha reliability coefficients for high school juniors and college freshman ranged from 0.86 to 0.92. With respect to validity, this scale correlated quite highly with the IPAT Anxiety scale and Manifest Anxiety scale; with correlation coefficients of 0.75 and 0.80 respectively (Dreger). This indicates that the scales measure basically the same construct.

The Computer Anxiety subscale of the Computer Attitude Scale consisted of 10 items and presented positively and negatively worded statements pertaining to feelings of anxiety towards computers. Subjects indicated which one of six ordered responses from

"Strongly agree" to "Strongly disagree" most closely represented the extent to which they agreed or disagreed with the ideas expressed. Scores could range from 10 to 60. Item responses were coded so that a higher score corresponded to a higher degree of computer anxiety. The alpha coefficient reliability of this subscale was reported by Loyd and Gressard (1984b) to be 0.86. The Attitude Questionnaire is found in Appendix C.

3.3.4 Survey of attitudes towards testing

This 15-item instrument was a modified version of Spielberger's (1977-80) Test Anxiety Inventory (TAI). It was used as a measure of the examinee's level of test anxiety while taking the science achievement test, and was administered to the paper-and-pencil test group. Five of the original 20 TAI items were omitted as inappropriate to the present study as they dealt with how the students felt about the test affecting their school performance. The remaining 15 items were revised to reflect situation-specific state test anxiety on the science test rather than general test anxiety. For example, "I feel confident and relaxed during tests" was changed to "I felt confident and relaxed during the test".

The items of the scale required a respondent to report on a variety of anxiety symptoms experienced during the test. Respondents answered in terms of a six-point Likert scale from "Strongly disagree" to "Strongly agree". Scores could range from 15 to 90 with higher scores indicating a higher level of test anxiety. Although the scale was a measure of test anxiety and the content of the items readily apparent, when presented to respondents it was called "Attitudes towards testing" as a precaution against cueing them that the instrument was an anxiety measure. The internal consistency reliability of this revised instrument as measured by the Hoyt estimate of reliability (Hoyt, 1941) was 0.93 for this sample of subjects. A copy of this questionnaire is found in Appendix D.

3.3.5 Survey of attitudes towards testing by computers

This instrument was administered only to the group tested by computer. It comprised four parts. In the first part, subjects were asked to indicate their age, sex, access to a computer at home, and amount of previous computer experience as measured by the number of hours spent on using a computer in a typical week, and the type of computer experience. The different categories classified under "type of computer experience" included "computerized games", "application programs" (such as wordprocessing, database and spreadsheets), "instructional programs" (such as tutorial, remedial and mastery learning), "computer programming", "courses other than computer programming courses", and "others". The second part consisted of the "Attitudes towards testing" scale described in the previous section 3.3.4.

The third part "Attitudes towards computer-based testing" was developed from the Computer Anxiety subscale (Loyd and Gressard, 1984b) described in Section 3.3.3. The 10 items were modified to reflect the respondent's level of computer anxiety while taking the computer-based science test. For example, "Working with a computer would make me very nervous" was changed to "Taking the test on a computer made me very nervous". The subjects placed themselves on a six-point attitude Likert scale anchored by the positions "strongly agree" and "strongly disagree". Scoring was "6" for "strongly agree" to "1" for "strongly disagree" in the case of statements expressing a negative effect, and was reversed for statements expressing a positive effect. Scores could range from 10 to 60. The Hoyt estimate of internal consistency reliability of the revised subscale was 0.87 for this sample.

The fourth part of the instrument consisted of open-ended free response questions about the examinee's reactions towards computer-based testing. Examinees were asked what they liked and disliked about testing by computers, if they would choose a computerized test, and reasons for their choice. Appendix E contains a sample of this questionnaire.

3.4 Research Design

A 2 X 2 factorial design was used as the experimental design for this study. The independent variables were:

1. Treatment (mode of test administration) having two levels i.e. computer-based test and conventional paper-and-pencil test.
2. Gender

The dependent variables were the achievement scores on the test and test anxiety scores.

The subjects in this study were 54 male and 51 female students selected from Grade 10 science classes in a secondary school in Burnaby. Within each class, subjects were blocked by gender and their scores on the previous school mid-term science exam (which was held a week prior to the study). Subjects were first separated by gender and rank-ordered by their exam scores. The top two boys were then randomly assigned to one of the two modes of test administrations: paper-and-pencil test or computer-based test. The next two boys were then randomly assigned to each of the two test treatments. This manner of random assignment was repeated until all the boys had been assigned. The girls were then assigned in a similar way. This procedure ensured an even balance of science abilities between the groups taking the two modes of test administration. The design can be diagrammed as shown in Table 2.

3.5 Experimental procedure and data collection

Subjects were told that participation in the study would entail approximately 45 minutes to 1 hour of their time, and that the study would take place during their science class period. Three days prior to the testing session, all students were given the "Attitude questionnaire" which was used as pre-measures of test anxiety and computer anxiety. They were then given the science achievement test in either the paper-and-pencil or computer-based format. Testing was conducted in groups of 20 to 23 students over 2 days on 5 occasions during the science periods.

Table 2 **2 X 2 (mode of test administration by gender) factorial design**

	Treatment (mode of test administration)		
Gender	Computer-based test	Paper-and-pencil test	Total
Male	27	27	54
Female	25	26	51
Total	52	53	105

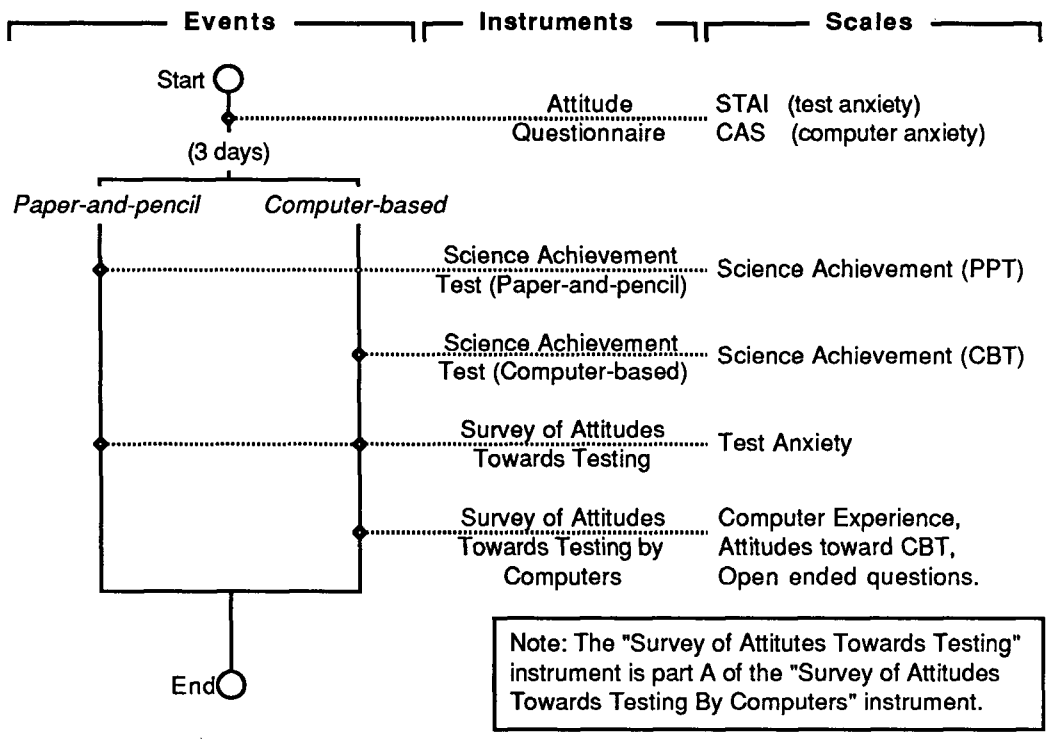
Students assigned to the computerized test were instructed to report to the school computer lab and given a brief 5 minute introduction to the use of the computer in taking the test. Each student worked on a separate computer. The computer presented the test instructions and the test items. The instructions were as similar as possible to those in the paper-and-pencil mode. Additional instructions which indicated which keys to use when taking the test on the computer were included. A copy of these instructions is found in Appendix A. The experimenter invigilated the computerized testing session and was present in the computer lab to clarify instructions pertaining to the operation of the computer, and to assist students having difficulty with the operation of the computer during testing. Students taking the paper-and-pencil test reported to a regular classroom in the same school building. They each received a test booklet containing test instructions and the 34 test items. The paper-and-pencil tests were invigilated by the science teacher. Examinees were not allowed to converse with each other during testing.

Upon completion of the test, students in the computer-based test group were administered the "Survey of attitudes toward testing by computer" questionnaire. Students in the paper-and-pencil group were given the "Survey of attitudes toward testing" questionnaire. A summary of the experimental procedure is shown in Figure 1.

3.6 **Analysis of data**

3.6.1 Preparation and coding of data

All demographic data, science test and attitude questionnaire item responses for each student were collected and transferred to computer data files. Item responses from the science test and each of the attitude scales were then scored using LERTAP (Laboratory of Educational Research Test Analysis Program: Nelson, 1974). Estimates of internal consistency reliability were also computed for each instrument. The composite data were used to provide a data base for subsequent analysis. The final data file contained four records per subject, each



Abbreviations used in the above table:

STAI : State-trait anxiety inventory

CAS : Computer anxiety subscale of the computer attitude scale

PPT : Paper-and-pencil test

CBT : Computer-based test

Figure 1 **Summary of experimental procedure**

headed by subject identification. The first record contained the demographic data, the second record the science scores, the third record the premeasures of test anxiety and computer anxiety, while the fourth record consisted of the test anxiety and computer anxiety scores. All statistical computations were performed using the SPSS-X program (SPSS Inc., 1988) on the Michigan Terminal System (MTS).

The four measures of science achievement, test anxiety, computer anxiety and previous computer experience were treated as continuous variables. The possible range of scores on these measures were 0 to 34, 15 to 90, 10 to 60, and 1 to 6 respectively. The derivation of the computer experience scale is explained in Section 3.6.3 below. The independent variables, mode of test administration (computer-based test; paper-and-pencil test) and sex (male; female) were treated as categorical variables.

3.6.2 Analysis of demographic data

The mean age and the percentage of subjects having access to a home computer were calculated. The proportion of subjects who checked each category of computer use and time spent on using a computer in a typical week was also determined.

3.6.3 Derivation of the computer experience scale

The computer experience scale was derived from a *post hoc* analysis of the responses to question 2 (time spent on using a computer in a typical week) and question 3 (type of computer use) of the "Survey of Attitudes towards testing by computers" questionnaire. First, the frequency of responses in each category of computer use time was noted. As there was only one response in each of the categories "15-19 hours" and "20 hours or more", the three categories "10-14 hours", "15-19 hours" and "20 hours or more" were combined to form one category "10 hours or more". Relative weightings of 0, 2, 4, 6 and 10 on a scale labelled USETIME were then respectively assigned to the categories "Never", "Occasionally", "1 to 4

hours", "5 to 9 hours" and "10 hours or more". The rationale for this was based on the assumption that if a student spent more time using a computer, he or she would gain a corresponding increase in the amount of experience that could contribute to a greater acquisition of computer skills.

Assignment of these weights was linked to the notion that if a student had never used a computer before, his or her experience would be virtually nil; and that the difference in computer experience gained by persons spending "5 to 9 hours" and "1 to 4 hours" a week on using a computer, is approximately the same as that of persons using it occasionally (once every few weeks) and persons never using a computer. Also, the increase in 4 points from a weight of 6 to a weight of 10, was based on the premise that the difference in the amount of computer experience acquired by students spending "10 hours or more" and "5 to 9 hours" is approximately twice the difference between users spending "5 to 9 hours" and "1 to 4 hours". Following from the above argument, a student spending "1 to 4 hours" a week on using a computer would be considered to have gained about twice as much experience as one using it only occasionally, and a student spending "10 hours or more" on using a computer would have gained more than twice the amount of experience than that acquired by a user spending "1 to 4 hours".

Next, the categories depicting the type of computer use were rank ordered according to the researcher's assessment of their relative difficulty and complexity. Weights of 1, 2, 3, 3, 5 and 7 were then respectively assigned to the categories "computerized games", "others", "instructional programs", "courses other than computer programming courses", "application programs" and "computer programming". This produced a scale with points labelled SUSE1 to SUSE6 representing each of the six categories. A composite scale ALLUSE based on an additive model and depicting all the categories of computer use checked by a respondent was then created by summing the 6 mutually exclusive subscales, SUSE1 to SUSE6.

The score on the ALLUSE scale represented the total score obtained by adding the scores on each of the 6 subscales. That is,

$$\text{ALLUSE} = \text{SUSE1} + \text{SUSE2} + \text{SUSE3} + \text{SUSE4} + \text{SUSE5} + \text{SUSE6}$$

Analysis of subjects' responses on the USETIME scale revealed that scores ranged from 0 to 10. On the ALLUSE scale, scores ranged from 0 to 15. A composite scale COMPUSE reflecting the amount of computer experience was then derived by combining the USETIME and ALLUSE scales in the following manner:

$$\text{COMPUSE} = (2 * \text{USETIME}) + \text{ALLUSE}$$

The USETIME scale was double-weighted for two reasons. Firstly, it seemed to be a more accurate measure of a subject's computer experience than the ALLUSE scale. A subject could have attained a high score on the ALLUSE scale by checking several categories of computer use even if he or she had encountered those uses only once. Hence by multiplying USETIME by a factor of 2, ALLUSE would not outweigh USETIME. Secondly, weighting USETIME by 2 allowed for a greater spread along the continuum of the COMPUSE scale. This composite scale had a range from 0 to 35 instead of 0 to 25 if USETIME had not been weighted by 2.

The final scale RTCUSE was obtained by a square root transformation of the COMPUSE scale.

$$\text{RTCUSE} = \sqrt{(\text{COMPUSE} + 1)}$$

This was done to produce a more normal distribution of scores as the scores distributed on the COMPUSE scale were positively skewed. Transformation of the scores to ensure normality was necessary as nonnormality would invalidate the standard tests of significance and the subsequent analytic methods employed since they were based on the normality assumption. The possible range of scores on the RTCUSE scale was from 1 to 6.

3.6.4 Science test analysis

Both the paper-and-pencil test and the computer-based test were scored with the number correct scale. An omission on an item was scored as an incorrect response. Test item analyses were then conducted for the paper-and-pencil and computer-based tests using LERTAP (Nelson, 1974). The means, standard deviations, Hoyt estimates of internal consistency reliability, and score distributions for both test modes were computed. The equivalence of the two test versions were determined by comparing these statistics for the computer-based test and the paper-and-pencil test.

The difficulty index (p-value) of each item, as measured by the percentage of respondents who got the item correct, as well as the distribution of the p-values of the items were compared for the paper-and-pencil and computer-based tests. The test of significance of the difference between proportions using independent samples (Glass & Stanley,1970 pp. 324-26) were used to test the significance of item difficulty differences between presentation media. This determined if the effect of mode of test administration was uniform over all science test items, which items were more difficult in the computer-based test mode, and which items were more difficult in the paper-and-pencil test mode. Any differences in item discrimination as measured by the point biserial coefficients were also analysed to determine which items, if any, contributed to overall test score differences. To determine if differences in item point biserial coefficients between the two modes of test administration were statistically significant, the test of significance of the difference between correlation coefficients using independent samples (Glass & Stanley, 1970, p. 311) was used. The distributions of scores obtained from both test versions were examined to see if both test administration modes produced normal distributions; they were compared using the Kolmogorov-Smirnov statistic and any differences between the empirical distribution of raw test scores tested for significance.

3.6.5 Analyses of science test, anxiety and computer experience measures

Pearson product moment correlations were computed to give an indication of the relationship among the scores obtained on the science test, test anxiety, computer anxiety and computer experience measures. Science achievement and test anxiety scores for the computer-based test and paper-and-pencil test groups were compared using analysis of variance (ANOVA) and analysis of covariance (ANCOVA) respectively. Mean score differences between groups on the science test and test anxiety scales were tested for significance at the $\alpha = 0.05$ level.

The first 3 research questions were whether the mode of test administration and gender result in differences in science achievement scores, and whether there is an interaction effect between mode of test administration and gender. To answer these questions, a two-factor fixed effects ANOVA with science scores as the dependent variable, and mode of test administration and sex as independent variables were used. The statistical hypotheses to be tested were :

H_{o1} : There is no difference between the mean science achievement scores of students tested by the computer-based test and paper-and-pencil test.

H_{o2} : There is no difference in science achievement scores between males and females tested by the computer-based test and paper-and-pencil test.

H_{o3} : There is no interaction effect between mode of test administration and gender.

Null hypotheses H_{o1} and H_{o2} represent the main effects of mode of test administration and gender respectively, whereas H_{o3} represents the interaction hypothesis.

To address the fifth research question of whether there is a difference in test anxiety level between the two groups taking the computer-based test and the paper-and-pencil test, an ANCOVA with the test anxiety scores as the dependent variable was used. The

independent variables were mode of test administration and sex; the scores on the premeasure of test anxiety were used as a covariate. An ANCOVA was performed to control for the effects of individual differences in general anxiety level. The statistical hypothesis tested was :

H_{o4} : There is no difference in test anxiety levels between students tested by the computer-based test and the paper-and-pencil test.

In order to determine the relationship between the amount of previous computer experience and computer anxiety level of students as a result of taking the computer-based test, regression analysis was used. The regression model used was :

$$Y = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

where Y is the adjusted computer anxiety score, the dependent variable

X_1 is the premeasure of computer anxiety on the "Attitudes towards computers" scale, the covariate

X_2 is the computer experience score on the RTCUSE scale

β_1 and β_2 are the regression coefficients associated with the covariate for computer anxiety X_1 and the independent variable X_2 respectively

ε is the error or residual term.

The scores on the "Attitudes towards computers" subscale, X_2 were used as a covariate. This was done to control for the students' computer anxiety level before taking the test. The parameter of interest was β_1. A significant and positive value would indicate a positive linear relationship between computer anxiety and computer experience, while a negative β_1 would indicate a negative relationship. If β_1 was found to be not statistically significant, no relationship between these two variables would be indicated.

To determine if the above proposed equation was a satisfactory model for analyzing the data, conventional tests were performed on the regression residuals (Norusis, 1983). These tests, which included the Durbin-Watson test (Neter & Wasserman, 1974) and analyses of residual plots, were conducted to check that the assumptions of regression analyses were not violated. They indicate respectively whether the errors are independent, random, and normally distributed with constant variance.

3.7 **Methodological assumptions**

The following assumptions were made when using the analysis techniques described in the aforementioned section.

1. The continuous variables are distributed normally and lie on an interval scale.
2. There is homogeneity of variance in the sampling distribution of the continuous variables.
3. All observations within the sample are independent.
4. The inherent assumptions of regression analysis (Pedhazur, 1982) which include :
   a) The independent variables are fixed. That is, if the experiment were to be replicated, the same values of the variable would be used.
   b) The independent variables are measured without error.
   c) The regression of the independent variable on the independent variable is linear.
   d) The model used is not mis-specified i.e. there is no inclusion of irrelevant variables nor omission of relevant variables from the regression equation, and the linear additive model postulated is appropriate.
   e) The errors are random, and normally and independently distributed with zero mean and constant variance.
5. There are no floor and ceiling effects associated with the science test. That is, the test is neither too difficult nor too easy that the distribution of scores is skewed.

Chapter 4 RESULTS

The results of the analyses described in the previous chapter are presented here. These analyses were conducted to answer each of the research questions posed in chapter 1. They evaluate the equivalence of the results obtained from the two modes of test administration, viz. computer-based and paper-and-pencil; compare the test anxiety level of examinees taking the two versions of the test; and determine the relationship between the amount of previous computer experience and computer anxiety of students taking the computer-based test. The psychometric characteristics of each of the instruments used are described and examinees' reactions to computer-based testing are also reported. The implications of the results of these analyses will be discussed in the next chapter.

4.1 Demographic information and students' uses of computers

The mean age of this sample of subjects was 15 years 9 months (standard deviation = 7.49 months). Of the students taking the computer-based test, 53.8 % reported having access to the use of a computer at home. Table 3 shows the distribution of students in each of the categories depicting the amount of time spent on using a computer in a typical week. A little more than a quarter (26.9 %) of the students said they did not use a computer, while 42.3 % used the computer for at least 1 to 4 hours in a typical week. The proportion of students checking each category of computer use is shown in Table 4. The most common uses of the computer were for playing computer games and application programs such as wordprocessing. The latter included writing essays and reports, and working on projects as part of schoolwork. Computer uses classified under "others" included telecommunications and experimenting with DOS.

Table 3 **Reported time spent by students on using a computer in a typical week**

Time	Percentage of respondents
Never	26.9
Occasionally (once in a few weeks)	30.8
1 - 4 hours	21.2
5 - 9 hours	9.6
10 - 14 hours	7.7
15 - 19 hours	1.9
20 hours or more	1.9

Note. n = 52

Table 4 **Proportion of students checking each category of computer use**

Computer use	Percentage of respondents
Playing computer games	44.2
Application programs	53.8
Instructional programs	19.2
Computer programming	5.8
Courses other than programming	9.6
Others	15.4

Note. n = 52

4.2 The computer experience scale

A single index of computer experience (RTCUSE) was constructed to represent an overall measure of a student's experience with computers. Constituent subscales included USETIME, indicating a measure of the amount of time spent on using a computer in a typical week, and ALLUSE, a measure of the level of sophistication of computer use. This scale was weighted in favour of computer experience offering more substantial exposure to the technology (for example, maximum weight was given to programming courses). The derivation of the scale was described in Section 3.6.3. Table 5 shows the descriptive statistics for the RTCUSE scale and its component subscales.

4.3 Comparisons of computer-based and paper-and-pencil science tests

The computer-based and paper-and-pencil tests were compared in terms of their means, dispersions, reliabilities, score distributions and item characteristics.

4.3.1 Psychometric characteristics

Table 6 shows the descriptive statistics for the two versions of the science test. The mean achievement score was 16.27 on the computer-based test and 13.75 on the paper-and-pencil test; the standard deviations were 4.54 and 5.33 respectively. No significant difference in variance of the science scores was found based on the test of homogeneity of variance using independent samples (Glass & Stanley, 1970, pp.303-306), $F(52, 51) = 1.38$, $p > .05$. Hence the standard deviations of the scores on both tests were not significantly different. The Hoyt internal consistency reliability estimate of 0.64 for the computer-based test was somewhat lower than the value of 0.75 obtained for the paper-and-pencil test. Standard errors of measurement of 2.67 for the computerized test and 2.63 for the paper-and-pencil tests were very similar.

Table 5 **Descriptive statistics for the computer experience scale**

Scale	Mean	S.D.	Min.	Max.
USETIME	3.2	3.1	0	10
ALLUSE	4.7	4.0	0	15
COMPUSE	11.1	9.0	0	35
RTCUSE	3.2	1.4	1	6

Table 6 **Reliabilities and descriptive statistics for the computer-based and paper-and-pencil versions of the science test**

	Mode of test administration	
	Computer-based	Paper-and-pencil
n	52	53
Mean	16.27	13.75
S.D.[a]	4.54	5.33
Maximum [b]	28.00	25.00
Minimum	7.00	5.00
Hoyt estimate of reliability	0.64	0.75
Standard error of measurement	2.67	2.63

Note.

[a] The difference between the standard deviations of the two test versions were not statistically significant.

[b] Maximum possible score on the science test = 34.

The difference in mean scores on the two versions of the science test was tested for significance using a 2-way ANOVA with mode of test administration and gender as main effects. Table 7 summarizes the results of this analysis. There was a significant difference for mode of test administration, $F(1, 101)=6.827$, $p < .05$ with the computer-based test group obtaining higher scores. No significant difference was found for gender as main effect. There was also no interaction effect between mode of test administration and gender; that is, the effects of the computer-based test and paper-and-pencil test were the same for both males and females. Mean scores and standard deviations for males and females are shown in Table 8.

4.3.2 Score distributions

The distributions of science scores obtained from both modes of test administration were tested for normality. Each of the score distributions was examined by the Kolmogorov-Smirnov one-sample goodness of fit test. As shown in Table 9, both the computer-based test and paper-and-pencil test produced normal distributions. The K-S Z statistic, and the associated probability values were $z = .586$, $p > .05$ and $z =. 753$, $p > .05$ respectively. Differences between the distribution of scores for the two test versions were tested using the Kolmogorov-Smirnov two-sample test. The results revealed a significant difference; $z = 1.522$, $p < .05$. Hence the score distributions on the computer-based test and paper-and-pencil test were unequal.

Figure 2 and Figure 3 show the distribution of science scores for the computer-based test and the paper-and-pencil test. The cumulative percentage distributions of scores for both modes of test administration are pictured in Figure 4. The curves do not cross, indicating that the cumulative percent of the number of examinees for the science scores on the paper-and-pencil test were consistently higher than for the computer-based test. The cumulative percentages of the number of examinees associated with the scores are found in Table F-1 in Appendix F.

Table 7 **ANOVA: Effects of mode of test administration and gender on science achievement scores**

Source of variation	SS	df	MS	F	Probability(F)
Main effects					
Mode of testing [a]	166.47	1	166.47	6.83	.010 *
Gender [b]	4.32	1	4.32	0.18	.675
2-way interaction					
Mode X Gender	61.02	1	61.02	2.50	.117
Residual	2462.71	101	24.38		
Total	2694.00	104	25.91		

Note. [a] n = 52 for computer-based test

n = 53 for paper-and-pencil test

[b] n = 54 for males

n = 51 for females

* $p < .05$

Table 8 **Means and standard deviations of science test scores for males and females**

Group	n	Mean	S.D.
Males	54	14.81	5.62
computer-based test	27	16.81	4.45
paper-and-pencil test	27	12.81	6.03
Females	51	15.20	4.50
computer-based test	25	15.68	4.66
paper-and-pencil test	26	14.73	4.39

Table 9 **Results of Kolmogorov-Smirnov tests for distribution of science scores**

	n	Mean	S.D.	K-S Z	Probability
One-sample test					
computer-based	52	16.27	4.54	.586	.882
paper-and-pencil	53	13.75	5.32	.753	.621
Two-sample test	-	-	-	1.522	.019 *

Note. Dashes indicate analysis is inapplicable.

* $p < .05$

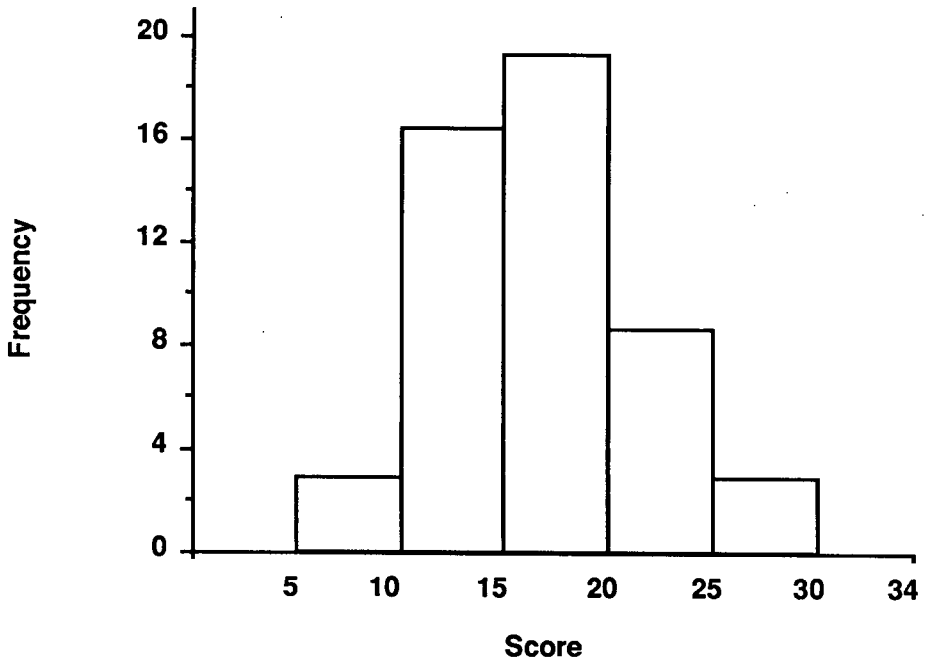


Figure 2 **Distribution of science scores for computer-based test**

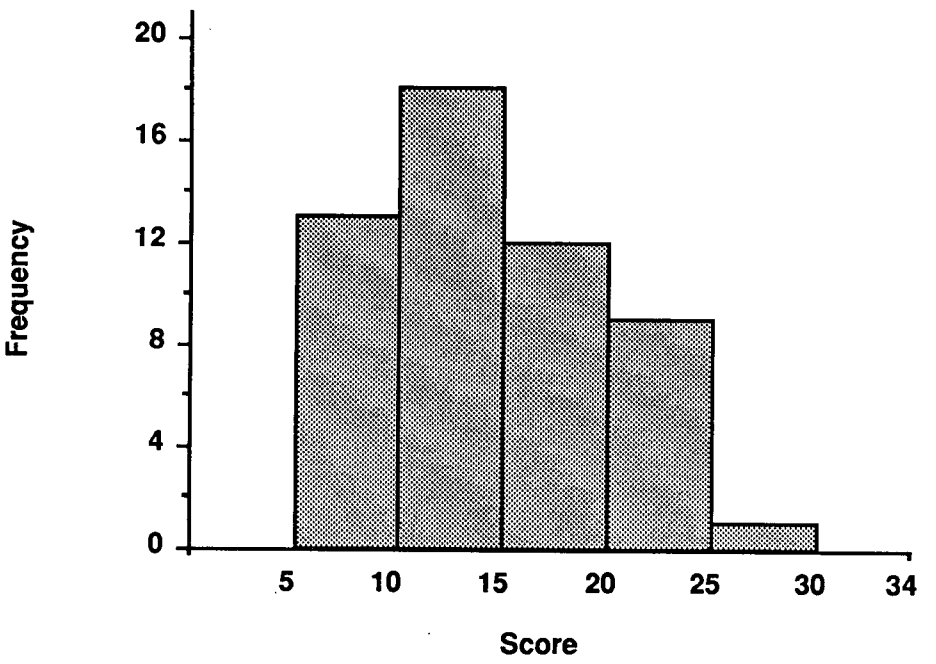


Figure 3 **Distribution of science scores for paper-and-pencil test**

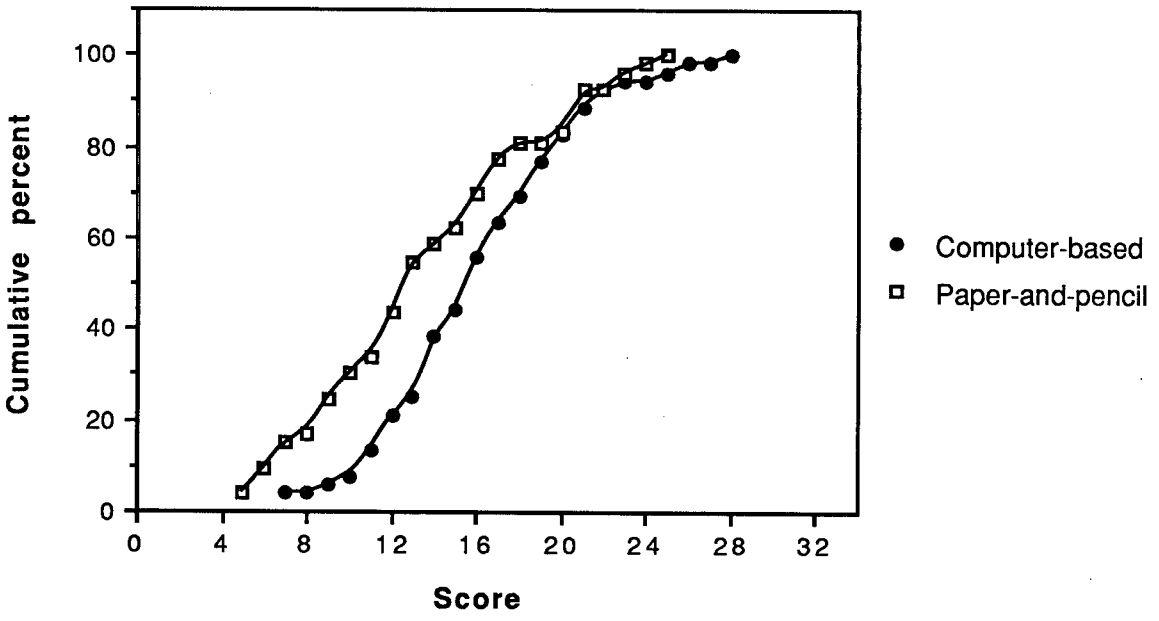


Figure 4 **Cumulative percentage curves of number of examinees for science test scores**

4.3.3 Item analyses: item difficulty and discrimination indices

Although the American Psychological Association's (APA) *Guidelines for computer-based tests and interpretations* (1986) do not specifically address differences in item characteristics (i.e. item difficulty and item discrimination) when comparing the equivalence between a conventional and computerized version of a test, item analyses were performed to test the similarity of item responses to both versions of the test, and to determine which items contributed to overall test score differences. The difficulty of an item was measured by the percentage of test-takers who got the item correct; that is by the difficulty index or p-value of classical test theory. Table 10 lists the p-values for all the 34 items on both versions of the science test. The p-values ranged from .23 to .81 for the computer-based test (mean = .48), and from .19 to .77 for the paper-and-pencil test (mean = .41).

To determine whether differences in item difficulties were statistically significant between modes of presentation, the test of significance of the difference between proportions using independent samples (Glass & Stanley, 1970, pp. 324-26) was used. Six of the 34 items had differences significant at $p < .05$; of these six, 3 were significantly different at $p < .01$. All these 6 items were more difficult i.e. they were associated with lower p-values, on the paper-and-pencil test. The items were questions 10, 12, 17, 27, 29 and 34 on the science test. The remaining 28 items were of approximately equivalent difficulty. The distributions of the p-values are shown diagrammatically in Figure 5 and Figure 6.

The item discrimination indices as measured by the point biserial coefficients on the two test versions are depicted in Table 11. An analysis of these coefficients by the test of significance of the difference between correlation coefficients using independent samples (Glass & Stanley, 1970, p. 311) revealed that the point biserial coefficients of only one item (question 8) were significantly different on the two test versions, $p < .05$.

Table 10 **Difficulty indices (p-values) for the two versions of the science test**

Item no.	Mode of test administration: computer-based [a]	Mode of test administration: paper-and-pencil [b]	z statistic
1	.40	.30	1.07
2	.31	.32	- .11
3	.33	.42	- .95
4	.48	.32	1.67
5	.35	.34	.11
6	.25	.40	- 1.64
7	.48	.40	.83
8	.33	.26	.79
9	.40	.38	.21
10	.54	.30	2.49 *
11	.23	.21	.21
12	.40	.19	2.36 *
13	.77	.62	1.67
14	.37	.42	- .52
15	.58	.51	.72
16	.48	.44	.41
17	.62	.32	3.08 **
18	.35	.40	- .53
19	.39	.42	- .31
20	.50	.40	1.03
21	.50	.40	1.03
22	.33	.36	- .32
23	.37	.30	.76
24	.48	.47	.10
25	.52	.45	.72
26	.67	.49	1.87
27	.71	.51	2.10 *
28	.73	.74	- .12
29	.81	.45	3.82 **
30	.39	.23	1.77
31	.35	.26	1.00
32	.75	.77	.24
33	.37	.43	- .63
34	.79	.55	2.61 **
	$\bar{p}$ = .48	$\bar{p}$ = .41	

Note. [a] n = 52
[b] n = 53
* $p < .05$
** $p < .01$

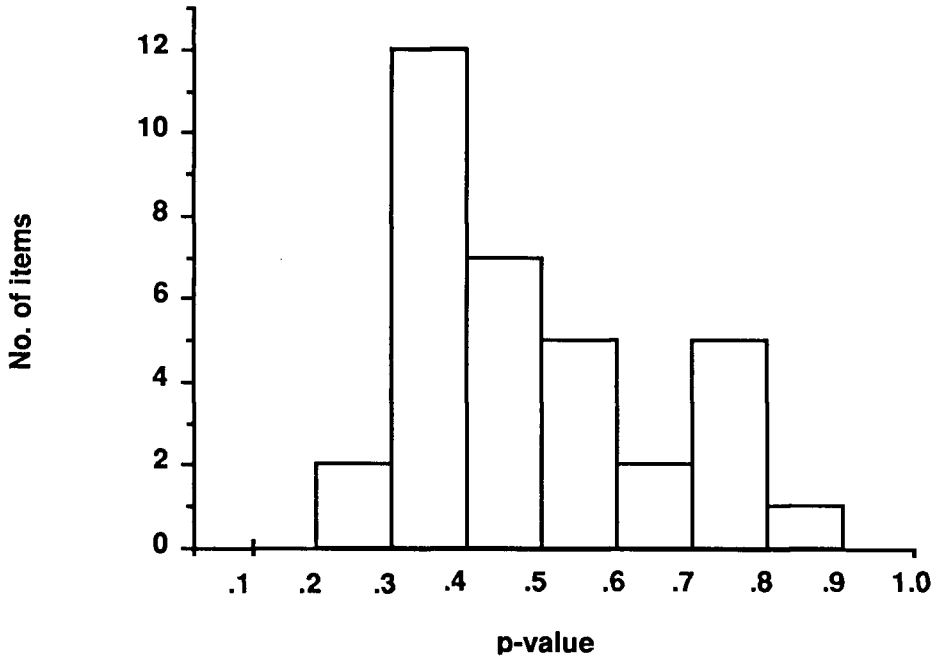


Figure 5 **Frequency distribution of p-values for computer-based test**

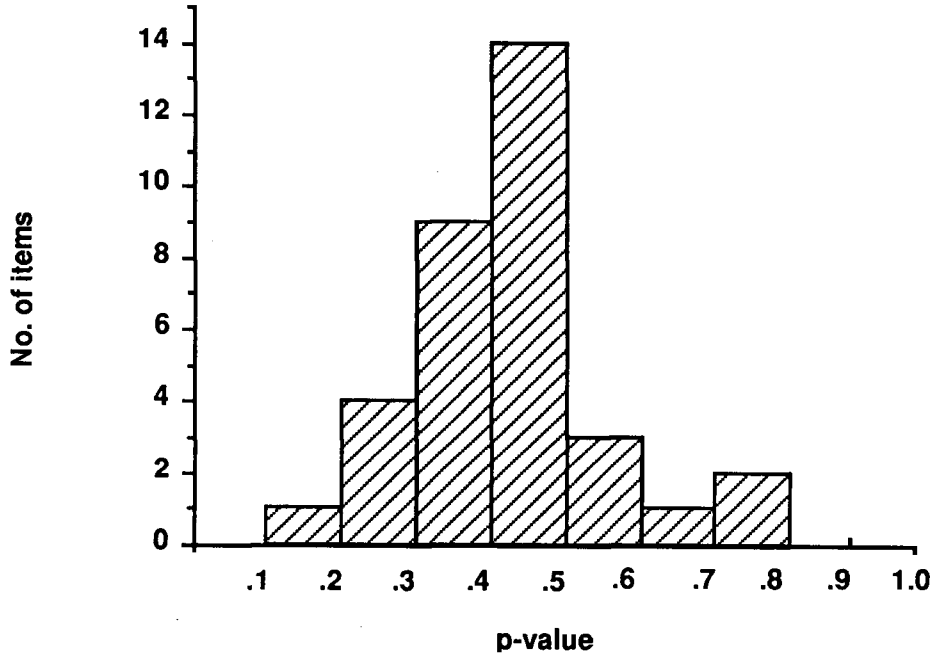


Figure 6 **Frequency distribution of p-values for paper-and-pencil test**

Table 11 **Point biserial coefficients for the computer-based and paper-and-pencil versions of the science test**

Item no.	Mode of test administration: computer-based [a]	Mode of test administration: paper-and-pencil [b]	z statistic
1	.27	.31	- .22
2	.15	.28	- .68
3	.28	.31	- .16
4	.34	.22	.65
5	.28	.09	.99
6	.32	.31	.05
7	.42	.63	- 1.46
8	-.06	.43	- 2.59 *
9	.25	.30	- .27
10	.12	.44	- 1.75
11	.22	.07	.77
12	.23	.33	- .54
13	.13	.37	- 1.28
14	.31	.37	- .33
15	.09	.21	- .61
16	.18	.35	- .91
17	.22	.25	- .15
18	.34	.26	.44
19	.24	.16	.42
20	.53	.43	.65
21	.23	.43	- 1.12
22	.38	.12	1.39
23	.44	.51	- .45
24	.48	.28	1.17
25	.37	.41	- .24
26	.52	.44	.52
27	.44	.54	- .66
28	.41	.19	1.21
29	.11	.44	- 1.80
30	.30	.25	.27
31	.41	.47	- .37
32	.21	.20	.05
33	.07	.33	- 1.36
34	.22	.38	- .88

Note. [a] $n = 52$
[b] $n = 53$
* $p < .05$

4.4 Psychometric properties of the affective scales

The descriptive statistics including the means, standard deviations and estimates of internal consistency reliability of the State-trait Anxiety Inventory (used as a premeasure of test anxiety) and Computer Attitude Scale (used as a premeasure of computer anxiety) for this sample of subjects are given in Table 12. Both instruments had a reliability of 0.83. The corresponding psychometric characteristics for the modified Test Anxiety Inventory and the Computer Attitude Scale are reported in Table 13. The reliabilities were 0.93 and 0.87 respectively.

4.5 Test anxiety

The mean test anxiety score was 35.55 for the computer-based test group and 40.73 for the paper-and-pencil test group. These results, together with the standard deviations and standard errors are presented in Table 14. As with the science achievement test scores, the Kolmogorov-Smirnov tests (Table 15) were performed to test the distributions of test anxiety scores for normality. They revealed that the distributions were normal, and not significantly different between the two modes of test administration.

To determine whether the difference in mean test anxiety scores was statistically significant, an ANCOVA was performed using the scores on the STAI (State-trait Anxiety Inventory) as a covariate. The results of this analysis are reported in Table 16. No significant difference in test anxiety level between the groups taking the two versions of the test nor between sexes was indicated. There was also no evidence of any interaction effect between mode of test administration and gender.

Table 12 **Reliabilities and descriptive statistics for premeasures of test anxiety and computer anxiety**

	Test anxiety STAI [a]	Computer anxiety CAS [b]
n	105	105
No. of items	20	10
Possible score range	20 - 80	10 - 60
Mean	41.33	22.39
S.D.	7.58	8.13
Maximum	73.00	49.00
Minimum	26.00	10.00
Hoyt estimate of reliability	0.83	0.83
Standard error of measurement	3.07	3.17

Note:. [a] Form X-2 of the State-trait anxiety inventory

[b] Computer anxiety subscale of the computer attitude scale.

Table 13 **Reliabilities and descriptive statistics for measures of test anxiety and computer anxiety**

	Test anxiety TAI	Computer anxiety CAS
n	105	52
No. of items	15	10
Possible score range	15 - 90	10 - 60
Mean	38.16	22.55
S.D.	13.61	8.52
Maximum	90.00	44.50
Minimum	15.00	10.00
Hoyt estimate of reliability	0.93	0.87
Standard error of measurement	3.40	2.88

Note. The Test Anxiety Inventory (TAI) and the Computer Attitude Scale (CAS) were modified to reflect examinees' level of test and computer anxiety while taking the science test.

Table 14 **Means and standard deviations of post-treatment test anxiety scores for computer-based and paper-and-pencil test groups**

Group	n	Mean	S.D.	Std. error
Computer-based test	52	35.55	11.76	1.631
Paper-and-pencil test	53	40.73	14.88	2.043

Table 15 **Results of Kolmogorov-Smirnov tests for distribution of post-treatment test anxiety scores**

	n	Mean	S.D.	K-S Z	Probability
One-sample test					
computer-based	52	35.55	11.76	.794	.554
paper-and-pencil	53	40.73	14.88	.551	.922
Two-sample test	-	-	-	1.324	.060

Note. Dashes indicate analysis is inapplicable.

Table 16 **ANCOVA: Effects of mode of test administration and gender on post-treatment test anxiety**

Source of variation	SS	df	MS	F	Probability(F)
Covariate					
Anxiety	766.73	1	766.73	4.40	.038
Main effects					
Mode of testing [a]	667.16	1	667.16	2.83	.053
Gender [b]	140.22	1	140.22	.81	.372
2-way interaction					
Mode X Gender	266.09	1	266.09	1.53	.219
Residual	17423.50	100	174.24		
Total	19261.25	104	185.20		

Note. [a] n = 52 for computer-based test

n = 53 for paper-and-pencil test

[b] n = 54 for males

n = 51 for females

4.6 Relationship between computer anxiety and previous computer experience

To assess whether previous computer experience was associated with the students' level of computer anxiety after taking the test on the computer, regression analysis was used with computer anxiety as the dependent variable, past computer experience as the independent variable, and the premeasure of computer anxiety as covariate to control for previous anxiety level. Table 17 contains the results of the analysis. While there was a significant relationship between students' prior computer experience and their computer anxiety before the test, no significant relationship existed between prior computer experience and the computer anxiety evoked as a result of taking the test ($F = .007$, $p >> .05$). Eighteen percent of the variance in computer anxiety could be accounted for by both computer experience and the premeasure of computer anxiety. However, most of this variance was explained by the premeasure of computer anxiety ($R^2 = .1795$) as after subtracting the proportion of variance accounted for by the premeasure of computer anxiety, computer experience accounted for only .01 % of the remaining variance in computer anxiety. This suggests that changes in computer anxiety that were apparently due to taking the test were not explained by computer experience.

An analysis of the regression residuals was conducted to determine if the linear regression model used was appropriate for the data, and to see if the model assumptions were met. No irregular trends were observed, indicating that there were no violations of the basic underlying assumptions of regression analysis, viz. that of normality, independence, homogeneity of variance, and linearity.

4.7 Intercorrelation among measures

The correlations among the measures of science achievement, test anxiety, computer anxiety and computer experience are given in Table 18. Science achievement was negatively correlated with test anxiety as well as the premeasures of test and computer anxiety. These correlations were low (ranging from -.18 to -.28) but statistically significant, $p < .05$.

Table 17 **Regression parameters and standard errors for the regression of computer anxiety on previous computer experience**

Variable	Coefficient	S.E.	F	Probability (F)
(constant)	11.8182	4.8257	5.998	.0179
premeasure of computer anxiety [a]	0.5128	0.1608	10.167	.0025
computer experience [b]	0.0683	0.8199	0.007	.9340
n = 52	R^2 = .1796	s = 7.8693 [c]		

Note.

[a] Proportion of variance in computer anxiety accounted for by premeasure of computer anxiety = .1795

[b] Additional proportion of variance in computer anxiety accounted for by computer experience = .1796 - .1795 = .0001

[c] s represents the standard error of R^2

Table 18 **Intercorrelations among measures of science achievement, test anxiety, computer anxiety and computer experience**

Measure	2	3	4	5 [a]	6 [a]
1. Science achievement	-.28 **	-.23 **	-.18 *	.18	.24 *
	(.002)	(.008)	(.031)	(.104)	(.044)
2. Pre test anxiety	- - -	.34 **	.20 *	.16	-.09
		(< .001)	(.021)	(.124)	(.256)
3. Pre computer anxiety		- - -	.25 **	.42 **	-.25 *
			(.005)	(.001)	(.035)
4. Test anxiety			- - -	.64 **	-.00
				(< .001)	(.493)
5. Computer anxiety [a]				- - -	-.10
					(.247)
6. Computer experience [a]					- - -

Note : N = 105

Below each correlation coefficient, the probability level is shown in parentheses.

[a] n = 52

* p < .05, one-tailed

** p < .01, one-tailed

There was a low positive correlation between science achievement and computer anxiety (r = .18); however, this was not statistically significant. All measures of anxiety were positively correlated with each other (r ranged from .16 to .64).

Of particular interest was the relationship between computer experience and computer anxiety. There was a significant negative relationship between the students' computer experience and their computer anxiety before the test (r = -.25). However, the correlation between computer experience and the computer anxiety evoked as a result of taking the test was negligible and statistically not significant (r = -.10). This finding is in agreement with the results of the regression analysis (described in section 4.6) that there is no relationship between computer experience and the computer anxiety of students after taking the computer-based test.

4.8 **Test-takers' reactions to computer-based testing**

Table 19 and Table 20 list test-takers' positive and negative responses to computer-based testing as well as the frequency of each type of response. Each student was asked to indicate what he or she liked and disliked most about testing by computer. Among the positive responses, the most popular reason for liking the computerized test involved the ease of answering questions and changing responses. Students appreciated not having to write, use a pencil or "bubble in" their answers on a scannable answer sheet. Other reasons mentioned were that the test seemed faster, was more interesting and was "fun".

There was a greater variety of negative responses. The most frequently cited response by the participants for not liking the computerized test was that it was difficult to review all answers at the end of the test and make changes as they had to move through consecutive items to go to a particular question. Students also said that they could not easily

Table 19 **Positive responses to computer-based testing**

	Response	Frequency
1.	Faster than written tests.	5
2.	More convenient, straightforward, simple to use.	8
3.	Easier to answer and correct a mistake.	4
4.	Test seemed easier.	12
5.	No writing involved -- just press a key.	16
6.	Neater.	3
7.	Feel comfortable.	3

Table 20 **Negative responses to computer-based testing**

	Response	Frequency
1.	Difficult to quickly review answers -- have to go through all consecutive questions.	5
2.	Difficult to skip questions and return to them later.	1
3.	Cannot see the whole test at a glance.	1
4.	Cannot draw on diagrams or "cross out" obviously incorrect answers.	1
5.	Don't like typing and using a keyboard.	3
6.	Pressing a wrong key may mean an incorrect answer.	2
7.	Computer makes too much noise when storing information on disk.	1
8.	Uncomfortable viewing -- bothers the eyes, headache, hard to read from screen.	3
9.	Fear of energy shutdown or computer ruining the test.	2
10.	Feel uneasy -- not used to it.	2
11.	Not private enough, harder to concentrate.	1

skip questions and come back to answer them later, and that they could not see the whole test at a glance. Some students noted that they did not like typing and having to use a keyboard, while others expressed that they felt uneasy working with a computer. Those in the latter category found it uncomfortable to read from a computer screen or were afraid of the computer "ruining the test". One student commented that he found it confusing to read from the computer screen and that it took him "double the time just to understand the question." In contrast to this, another student who had indicated a liking for computer-based testing said that he "liked reading from a computer screen as it made the test seem easier." A study of Table 21 reveals that the negative responses can be classified under 3 general categories. Responses 1 to 4 are related to the flexibility of the test-taking task, responses 5 to 8 involve computer-linked factors, while responses 9 to 11 pertain to human factors.

When asked if they would choose a computer-based test over a paper-and-pencil test, 71.2 % of the 52 students who had taken the computerized test indicated that they would. The most common reason given for this preference was that the test was "easier"; 16 of the 37 students who opted for the computerized test gave this reason for their choice. Other reasons included the following:

a) "The test took less time."

b) "I felt better / more relaxed / more comfortable."

c) "I didn't have to write, erase or scratch out wrong answers..... just pressed a key. It was not a hassle."

d) "I'd rather type than write or fill in bubbles."

e) "It was easier to think about the answers."

Of those not stating a preference for computer-based testing, the most common reasons were that they did not like computers, and that it was much easier to read on paper because they were able to see all the questions and responses simultaneously. Some indicated

that they were familiar with paper-and-pencil tests and were resistant to change. Students' comments included "I prefer a paper-and-pencil test because when I am finished with a test, I like to be able to glance over everything to see if it's the way I like it", and "thinking about computers making mistakes while I was doing the test kind of bothered me."

4.9 Summary

The computer-based and paper-and-pencil science tests were compared with regard to their means, variances, score distributions, reliabilities and item characteristics. The means for the two versions of the test were significantly different, with the group taking the computer-based test obtaining higher scores. Differences in mean scores between sexes were not significant; there was also no interaction effect between mode of test administration and gender. The variances of the science scores were similar for both modes of test administration. However, the score distributions were significantly different. The estimate of internal consistency reliability was 0.64 for the computer-based test, and 0.75 for the paper-and-pencil test. Six of the 34 items had significantly different p-values; all were lower on the paper-and-pencil test. Only 1 item had a significantly different point biserial coefficient on both versions of the test.

There was no significant difference in test anxiety level between the groups taking the computer-based test and the paper-and-pencil test. While there was a significant relationship between students' previous computer experience and their computer anxiety before the test, no such relationship existed between computer experience and the computer anxiety evoked as a result of taking the test. Any changes in computer anxiety due to taking the test were not explained by computer experience. Overall, 71.2 % of the students who took the computer-based test indicated that if given a choice, they would prefer to take the test on a computer. The most frequently cited reason for this was that the computer-based test seemed "easier".

Chapter 5 DISCUSSION, CONCLUSIONS AND RECOMMENDATIONS

5.1 Summary of the study

The aim of this research study was to compare the effects of test presentation on students' performance and test anxiety level under two modes of test administration; viz. computer-based test and conventional paper-and-pencil test, in a secondary school classroom setting. The study sought to determine whether the mode of test administration, and gender resulted in differences in achievement scores and test anxiety level, and whether there was an interaction effect between mode of test administration and gender. The possible effect of gender was of interest as sex differences in computer usage and attitudes towards computers have been indicated, (for example, Chen, 1986) with males being more familiar with, and having more positive attitudes towards computers. The study also assessed the equivalence of the computer-based and paper-and-pencil tests in terms of achievement scores and item characteristics, explored the relationship between computer anxiety and previous computer experience, and investigated the affective impact of computerized testing procedures on students.

A 2 X 2 (mode of test administration by gender) factorial design was used. A sample of 105 Grade 10 students participated in the study. Subjects of each gender were randomly assigned to take either a computerized or paper-and-pencil version of a science test. Three days before taking the test, all students were given the "Attitude questionnaire", which included premeasures of test and computer anxiety. The students then took the science test in either the computer-based or paper-and-pencil format. To enhance the internal validity of the study, both test versions were designed to control extraneous influences by using identical test time limits, number of questions, item content and sequence; all items were multiple choice questions. The item administration procedures were also made as identical as possible for each presentation medium. The computer-based version was written to mimic as closely as possible the flexibility of the paper-and-pencil test format. Test-takers were able to move back and

forth through the test, review previous responses, and change answers. Immediately after the test, students in the computer-based test group were asked to complete the "Survey of attitudes towards testing by computers". This consisted of questions designed to determine the participants' previous experience with computers, their test anxiety and computer anxiety level while taking the test, and their reactions towards computer-based testing. Students in the paper-and-pencil test group answered the "Survey of attitudes towards testing" questionnaire designed to measure their test anxiety level while taking the paper-and-pencil test.

The results of the study indicate that achievement scores on the science test were significantly higher for the group taking the computer-based test ($p < .05$). The score distributions for the two groups were also significantly different. Although both the computer-based test and the paper-and-pencil test produced normal distributions, the mean score for the computer-based test was higher. However, the variances of the test scores were similar for both versions of the test. No significant difference between sexes in mean scores was observed; there was also no interaction effect between mode of test administration and gender. The internal consistency reliability estimate was 0.64 for the computer-based test and 0.75 for the paper-and-pencil test. Difficulty indices or p-values for 6 of the 34 items were significantly higher on the computer-based test; that is, all these 6 items were easier on the computerized test. The item discrimination index, as measured by the point biserial correlation coefficient, differed significantly between the two test versions for only one item.

No significant difference in test anxiety level was observed between the groups taking the tests under the two modes of test administration. A significant relationship existed between students' prior computer experience and their computer anxiety before the test. However, there was no significant relationship between their previous computer experience and the computer anxiety evoked as a result of taking the test on the computer. The change in computer anxiety that was due to taking the test was not explained by computer experience.

When given a choice between a computerized test and a paper-and-pencil test, 71.2 % of the students who took the computer-based test indicated that they would prefer to take the test on a computer. A little less than half of these students said that they found the computerized test "easier". In response to what they liked most about computer-based testing, students indicated that they felt that the test took a shorter time and was more convenient and easier to answer than a conventional paper-and-pencil test. They also appreciated not having to write, erase mistakes or fill in bubbles on a scannable sheet. Responses to what students disliked about computerized testing included the difficulty involved in reviewing and changing answers, having to type and use a keyboard, fear of the computer making mistakes, and a feeling of uneasiness because the medium of item presentation was different from what they were accustomed to.

5.2 **Discussion and implications of the research findings**

5.2.1 Students' performance on the science test

The results of this study indicate that the computer-based and paper-and-pencil test versions of the science test were not equivalent. Despite being identical in length, item content and sequence, they did not meet all the criteria for score equivalence as established by the American Psychological Association's (APA) *Guidelines for computer-based tests and interpretations (1986).* Although the dispersions of the score distributions were approximately the same, the distributions of the test scores on the two test versions were significantly different. Furthermore, the difference in mean scores were statistically significant for the tests given under the two modes of test administration, with the computer-based test having a higher mean score. Contrary to most of the earlier studies on the comparability of computer-based and paper-and-pencil tests, the subjects in this study achieved significantly higher scores on the computerized test than on the paper-and-pencil test. In previous research, several studies (for example, Eaves & Smith, 1986; Olsen *et al*, 1986; Spray *et al*, 1989; Wise & Wise, 1987) found score differences between test modes to be non-significant. Where significant differences were observed, the mean scores on tests administered

by computer were usually lower (for example, Lee, 1986). Almost all of these previous studies however, focused on adult populations. The only exceptions were the investigations by Olsen *et al* (1986) and Wise & Wise (1987) which were conducted on elementary school children.

The finding that the mean achievement score on the computer-based test was higher than that on the paper-and-pencil test was unanticipated. Because the two test versions were identical in length, item content and sequence, it was hypothesized that the test score means would be similar. If a difference in mean test scores were indicated, it was expected that scores on the computer-based test would be lower. This was because of the limited flexibility in reviewing responses on the computer-based test, as well as the possible debilitating effects of increased test anxiety and computer anxiety associated with taking a test administered in an unconventional manner. However, these results were not obtained. The experiential background of the subjects may have played a part in their ability to respond favorably to the computer-based test. It would seem reasonable to assume that high school children would be more likely than elementary school children to have gained the experience that would allow them to adapt to the unfamiliarity of using computers in testing. Morever, when compared with older adults, it also seems plausible that the teenagers of today are less likely to be affected by the adverse effects of computer anxiety at a time when the availability of computers is more commonplace, and when many students tend to become familiar with computers at a young age.

One possible explanation for the higher achievement scores obtained on the computer-based test was that students taking this version of the test tried harder. An analysis of students' responses to each of the test items revealed that there was a tendency for test-takers on the computerized test to avoid option E, "I don't know". For 28 of the 34 items, the proportion of students selecting option E was lower for the computer-based test than for the paper-and-pencil test. Also, the average number of E's selected by a test-taker was 3.9 for the

paper-and-pencil test, but only 1.3 for the computer-based test. In other words, this option E tended to be selected far less often on the computerized test than on the paper-and-pencil test. Hence by narrowing their choices to options A to D, there was a greater probability of getting a correct answer. This interpretation is supported by the finding that for most of the test items, the p-value or proportion of correct answers on the computer-based test was higher. An implication of this finding is that even though the content of the items is the same, mode of presentation can make a difference in the propensity to guess on an item. This difference in test-taking behaviour has also been suggested by Greaud & Green (1986) as a possible explanation for score differences between computerized and paper-and-pencil tests.

A plausible reason for the greater tendency for students to guess on the computer-based test may be that the computerized mode of test administration forced them to concentrate harder on each question, presumably because only one question was in view at any one time. On the computerized test, each question was presented individually, whereas two items were printed per page on the paper-and-pencil test. This modification in the way items were presented could have affected test-taking behaviour. Another reason for students trying harder and consequently tending to avoid the option E "I don't know" may be attributable to the novelty effect of computer testing. This might have served as a motivational factor to heighten their attention and keep them on task . Yet a third reason may be that students concentrated harder on a question simply because they may have found it more difficult to read from a computer screen than from a piece of printed paper.

The finding that the mean achievement score on the computer-based test was not lower than that on the paper-and-pencil test may indicate that if the tasks involved in taking a test on the computer are kept simple enough, even test-takers with minimal computer experience will not be disadvantaged by having to use an unfamiliar machine. The computer-based test was designed to be easy to use, and to avoid giving users with extensive computer

experience an advantage over novice users. An interesting question to consider is whether more complex response demands used in a microcomputer testing program might lead to different results. For example, if test-takers were required to make use of several keystrokes in response to each stimulus item, or type a short answer or sentence instead of simply selecting one of the options "A" to "E", then a special advantage might be accorded to students with relatively large amounts of computer exposure.

The estimate of internal consistency reliability was slightly lower for the computer-based test than the paper-and-pencil test. This indicates that measurements of students' performance on the test items obtained on the computerized test were not as consistent as those obtained on the paper-and-pencil test. The differences in manner of presentation of test items and test-taking flexibility might have elicited a change in the strategies with which students approached the items. For example, on the computer-based test it was difficult, though not impossible, to review and revise an earlier response. There was also a certain inherent delay in retracing the test to earlier items. In addition, the opportunity to obtain cues from other items was limited as the items were presented individually. It is possible that these factors served to decrease the reliability of the computer-based test.

That students' responses could have been influenced by their opportunity to obtain cues from other items raises an interesting question. Which administration mode yields scores that are less influenced by test-taking ability and that more accurately reflect a student's true mastery and knowledge of the subject matter? In a computer-based test, it is harder to gather content-based cues and assimilate information from other items. However, for a paper-and-pencil test, such cross-item inference is much easier. Consequently, if deductive reasoning strategies are employed when answering a question, it is more likely that a test-taker on a paper-and-pencil test would be able to gain points beyond that which he or she would receive on the basis of sure knowledge of the specific subject matter. In these strategies, which

deal with methods of obtaining the correct answer indirectly or with only part of the knowledge necessary to answer a question, the correct answer itself would not be known if no choices were given, or if no other questions were asked.

As mentioned earlier, the results showed that the item difficulty indices were affected by mode of test administration, with higher values associated more frequently with the computer-based test. This effect was fairly uniform across the items. Where there was a significant difference in difficulty index between the two versions of the test, the item on the computer-based test was consistently easier than on the paper-and-pencil test. All the 6 items that had significantly different difficulty indices were reviewed to determine whether any surface features of the item could be identified that could explain the observed results. The features considered included the length of the question stem and response options, the use of diagrams, as well as the topics and domains tested. The items examined were questions 10, 12, 17, 27, 29 and 34 on the science test. In all of these items, the question stems and response options were of average length. No diagrams were used in any of the questions. Two of the questions were from the "Physical sciences" (items 10 and 34), two from the "Life sciences" (items 12 and 27), and the remaining two from the "Earth / Space sciences" (items 17 and 29). The domains tested included "Skills and processes"(items 12 and 34), "Knowledge" (items 17, 27 and 29), and "Application of science concepts" (item 10). The fourth domain, "Rational and critical thinking" was not associated with these 6 items. Hence, these 6 items covered all the 3 topics tested, and all but one of the 4 domains. This review identified no distinguishing features that could be used to explain the differences observed.

5.2.2 Relationships among test anxiety, computer anxiety, and computer experience

The mean *test anxiety* score was 35.55 for the computer-based test and 40.73 for the paper-and-pencil test on a scale where the possible range of scores was from 15 to 90. This reflected relatively low levels of test anxiety. A possible explanation for this is that the

students knew that their scores on this test would not affect their overall school science grade. Hence their "true" test anxiety may not have been evoked during the test. The test anxiety scores associated with the two versions of the test were not significantly different. These results indicating no differences in test anxiety level between the groups were somewhat unexpected; they are also contrary to the findings by Llabre *et al* (1987) and Ward *et al* (1989) which showed that college students taking a computer-based test reported higher levels of test anxiety than their counterparts taking a paper-and-pencil version of the same test. Apparently, the new and unfamiliar experience of taking a test on a computer did not raise the test anxiety level of the students. A plausible reason for this may be that many of the students taking the computer-based test were already familiar with the use of a computer; consequently the thought of having to take a test on the computer did not arouse much test anxiety. 26.9 % of the students in the current study never or hardly used a computer. In contrast, in Llabre *et al's* (1987) study cited above, 77.5 % of the subjects reported never or seldom using a computer.

That students had few qualms about taking the test on the computer is attested to by the fact that the students in the computer-based test group as a whole reported fairly low levels of *computer anxiety*. The average computer anxiety score was 22.55, below the mid-point score of 35, on a 10 to 60 point scale. The observed low computer anxiety may be a reflection of the students' prior exposure to computers. It is noted that more than half of the students (53.8 %) had access to a home computer. Furthermore, the school which participated in the study had a comparatively large Macintosh lab of 16 microcomputers, so the students had routine access to the use of computers. Although microcomputers were not used in the science curriculum, some of the students used the computer regularly for working on their English essays and projects in other classes. 30.8 % of the students used the computer once in a few weeks, while 42.3 % used it for at least 1 to 4 hours in a typical week. In all, 73.1 % of the students made use of a computer on a regular basis. Only 26.9 % of the students reported not using a computer.

It appears then, that the findings in this study may have been related to certain aspects of the different population studied. The Grade 10 students in the present study differed from the subjects in the previous studies by Llabre *et al* (1987) and Ward *et al* (1989) both in terms of their age and prior computer exposure. Compared to the subjects in both these studies, the students in this study were younger and had more computer experience prior to the investigation. The latter finding may be due partly to the fact that computers are more easily available nowadays than they were several years ago, and that the students today have more opportunities to become familiar with computers at an early age. It is likely therefore, that the students would have felt less anxious when using a computer, and were better able to adapt to the novelty of a computer-based test. These differences may be important because it suggests that the attitudes of high school students toward computer-based testing may not be the same as those of subjects who have less prior computer experience, or who are older and perhaps less likely to adapt to the unfamiliar mode of computerized test administration.

The results of this study also showed that computer experience had little, if any impact on the students when they were taking the test. There was no significant relationship between the students' prior computer experience and their computer anxiety as a result of taking the test, after controlling for their computer anxiety before the test. It may be that the computerized test-taking tasks were simple enough so as not to elicit much computer anxiety, even in students with little prior computer experience. If this were indeed the case, then the results of the current study imply that if the procedures involved in taking a computer-based test, such as the keystrokes and commands required to view text and to record responses, are kept uncomplicated, then individual differences in terms of previous experience with computers may not influence computer anxiety when testing high school students. It is also encouraging to note the lack of a significant relationship between computer experience and computer anxiety as a result of taking a computer-based test, as this indicates that the computer-based testing

experience did not arouse increased computer anxiety among students with minimal exposure to computers.

5.2.3 Correlates of test anxiety, computer anxiety and computer experience with achievement

Although the relationships between test anxiety and achievement, computer anxiety and achievement, and computer experience and achievement were not the main questions of interest in this study, they were examined as part of a *post hoc* analysis. It must be borne in mind however, that no explicit causal statements should be made regarding the relationship between the measures. The relationship between these measures was calculated using Pearson's product moment correlation; a premeasure of science ability had not been used as a control to remove the effect of different science abilities. Nevertheless, this correlation does, to some extent, indicate the degree of association between the measures of interest.

The results of the correlational investigations reveal that the relationships between the above-mentioned measures were low. The correlation between test anxiety and achievement was - .18 ($p < .05$). This suggests that an increased test anxiety level had a slightly deleterious effect on achievement, an observation which is supported by most of the research literature. Although a significant inverse relationship between the premeasure of computer anxiety and achievement on the computer-based test was indicated ($r = -.23$), the correlation between students' achievement and their computer anxiety after taking the test was positive but not significant ($r = .18$). It appears then that the computer anxiety experienced by the subjects as a result of taking the computer-based test was not associated with a lowered achievement level. As noted earlier, the average computer anxiety was low for the group as a whole. Any computer anxiety aroused by having taken the computer-based test may thus not have been sufficiently debilitating as to depress test performance. For some individuals, this increase in anxiety may even have produced a facilitating effect and resulted in improved

performance. This may explain the slightly positive direction of the effect of computer anxiety upon achievement.

Previous computer experience was positively related to achievement scores (r = .24), with more experience associated with higher scores. This result however, should be interpreted with some caution; it does not necessarily mean that there exists a direct *causal* link between them. This need for caution is probably best expressed in the oft-repeated admonition "Correlation is no proof of causation" (Pedhazur, 1982, p. 579). The correlation between these two variables is low. The correlation approach is not sufficiently powerful and is a "potentially misleading test for causation when used alone" (Glass & Hopkins, 1984, p. 104). Variables other than the two under consideration, such as cognitive ability and home background could contribute to the observed association; the relationships that exist among all these variables may be too complex to be explained in terms of a single cause. Part of the reason for the observed positive, albeit low correlation between computer experience and achievement may be that students who are higher achievers (and hence who would have scored higher on the science test) come from more advantaged backgrounds where the opportunity for access to computers is greater.

5.2.4 Students' reactions to computer-based testing

Part of the responses elicited by the questionnaire dealt with students' written comments about computerized testing. The students generally had a positive attitude toward computer-based testing. Almost three quarters (71.2 %) of the students indicated a preference for computer-based tests over paper-and-pencil tests. Testing by computers was perceived by about one-third of the students as "easier" than conventional testing by paper-and-pencil. This can be interpreted in three ways. It can mean that students found the procedure involved in keying in and changing their answers easier, as pressing a key was faster, neater and more convenient than having to write, use a pencil, or erase wrong answers. The second possible

interpretation is that because the questions were presented individually, it was easier to concentrate on each question as no others were in view. Yet a third interpretation is that the students actually found the test less difficult. This is in agreement with the finding that the overall mean p-value of the computer-based test was higher than that for the paper-and-pencil test.

It was also interesting to note that some students expressed that they felt more confident taking a test on computer than by conventional methods, and that they actually found taking the computer-based test more "comfortable" and "relaxing". Perhaps for some individuals, especially those with extensive computer experience, taking a computerized test may actually be a more positive and less stressful experience than taking a paper-and-pencil test. The test-taking procedures used in this study were rather straightforward. However, if the test had employed a different item format, where more complicated response demands such as the use of more keystrokes, were expected of the students, then the attitudes elicited might not have been so positive. Students with little computer experience might then find the procedures involved confusing, or even overwhelming. It is possible also, that with questions tapping into cognitively more demanding domains, such as that requiring problem-solving skills, students may actually find a paper-and-pencil test easier to take, as they can sketch diagrams, manipulate symbols, or write on paper rather than depend only on mental imagery to arrive at their answers.

The most common problems encountered by the students involved the difficulty with which earlier items could be reconsidered. To review their answers on the computer, the students had to retrace previous questions consecutively. The students' remarks revealed that procedural variables in the administration of computerized tests are important determiners of the observed affective reactions. Because the method of administration of computer-based testing is different from conventional testing, care should be taken to ensure that the test-

taking flexibility, devices used and the environment in which the computerized test is taken is conducive to optimal test performance.

Most of the students did not find the experience of taking a test on a computer intimidating. Of the 52 students who took the test, 3 students however showed strong negative reactions towards computer-based testing. One expressed an intense dislike for computers while the other two indicated that they were familiar with paper-and-pencil tests and were resistant to change. One of the students even commented that "the school board should continue with paper-and-pencil tests", and that he felt he would have done much better had he taken the test using paper-and-pencil. He went on further to say that "teachers have been doing it this way (that is, administering tests by paper-and-pencil) for centuries, so why stop now? Just because it's out of date?". Even though these sentiments were expressed by only a few individuals, they are important and should not be dismissed when evaluating the affective impact of computer-based testing. The negative feelings experienced by students such as these, may prevent them from functioning well on a computer-based test. Two of the students cited above reported computer anxiety levels above that of the group average. One of the students had relatively little computer experience. This finding suggests the need to familiarize test-takers with the technology used in computer-based tests prior to test administration. Perhaps the negative attitudes of some students toward computerized testing would become more favorable as the students become acclimated to the machinery and its associated procedures.

5.3 Limitations

Although extraneous variables that could influence the dependent measures were controlled for, both by design and the use of randomization, a few factors that could have affected the sensitivity of the study will be discussed.

An important limitation involved the subjects' level of motivation and test anxiety. Because of constraints imposed by the university ethical review committee regulations, students were told that their scores obtained on the science achievement test would not be included in their overall school science grade. Consequently, their test anxiety aroused in taking the science test was low. A higher level of test anxiety might have been evoked in the students, had the test scores been included as part of the school science grade. To minimize this effect, students were told that although their scores would not contribute to their overall grade, they would be given to their science teacher for purposes of instructional planning.

Although the test-taking flexibility of both versions of the science test were made as similar as possible, there was one difference in test-taking flexibility. In reviewing their answers, subjects taking the computer-based test had to proceed question by question. They had to go either to the question immediately preceeding or following the one that they were working on, and did not have the freedom to turn directly to a particular question. The subjects in the computer-based test group knew they were participating in an experiment and experienced the novelty of it, so they may have tried harder (Hawthorne effect). The physical setting of the classroom and computer lab were not identical. Because of the keystrokes involved in using the keyboard, students in the computer-based test group experienced a slightly higher noise level than those in the regular classroom. This may to a certain extent, have affected the students' concentration and performance.

The results of this study have some, though limited generalizability to the target population of secondary school students. They may not generalize to other grade levels, subjects, schools where students come from a different socio-economic background, or schools where access to computer facilities is different. Also, in this study, the test items were arranged in descending order of difficulty. Since this order of item presentation was different

from convention, these results may not generalize to a test where the ordering of the test items is from easy to hard.

5.4 Suggestions for future research

As the technology for computerized test administration becomes more accessible, there are possibilities for future related research in many areas. To facilitate discussion, these areas are divided into three groups. Clearly, these groups are not mutually exclusive; they serve only as a means of identifying the focus of each group of suggested studies.

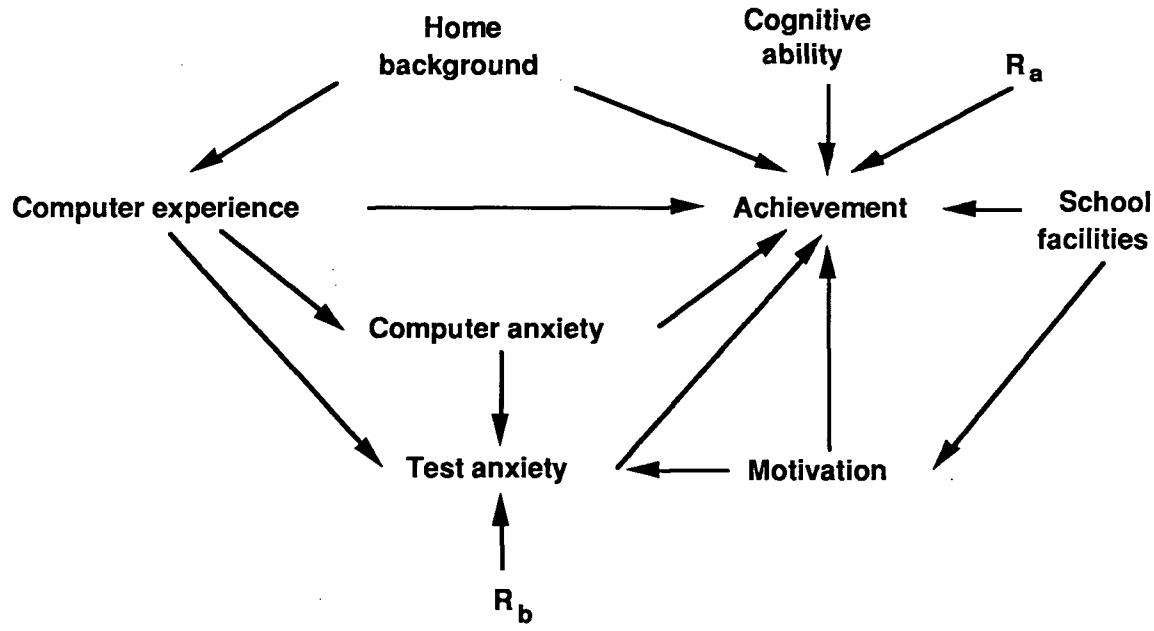
The first group of studies would address the generalizability of this study's results to other populations and achievement tests in different subject areas. For example, the study could be replicated with students of a younger age group, such as elementary school students who are at a different developmental stage and have different amounts of computer experience. Replication with a test in a different subject area may answer pertinent questions such as: would the results differ across a range of subjects? Some students may perceive that computerized tests are more appropriate for certain subjects. In fact, one of the students in this study commented that "tests in some subjects such as Math and English, would work well on computers". This association of computer-based tests with specific subject areas may be a result of the student's prior experience with using computers to write essays in English classes, or to do Mathematics remedial exercises.

The next group of suggested studies stems from the results and implications of the current study. Because differences existed between the computer-based test and paper-and-pencil test in test-taking flexibility, opportunity to obtain cues, and tendency to guess on a question, further studies could be carried out to compare the test-taking behaviour for these two modes of test administration, as measured by the number and pattern of changed answers. The relationship between the number of times the response to an item is changed and the

difficulty index of an item can also be explored. The answers to these questions may give an insight into response style analysis in computer-based testing. As the computer is capable of recording the time taken to answer each question, a fine analysis of performance and response latency is also possible. Research could be done to evaluate the additional information that can be obtained from individual item responses.

Based on the variables relevant to the present study, the model shown in Figure 6 is proposed to depict some of the inter-relationships among computer experience, computer anxiety, test anxiety and achievement on a computer-based test, and their hypothesized directional effects. It is a simplified version of one of various conceivable models designed to portray the possible pattern of causal relations among the above set of variables. Paths, or posited linkages between the variables of interest, in the form of unidirectional arrows, are drawn from the variables taken as causes (independent variables) to the variables taken as effects (dependent variables). In the model, computer experience for example, is postulated to exert both a direct as well as indirect effects on achievement. That is, part of its effect is seen as mediated or transmitted by intervening variables such as computer anxiety.

Besides the 3 variables pertinent to the current study, viz. test anxiety, computer anxiety, and computer experience, a host of other inextricably linked factors are related to student achievement as well (Atkinson, 1974; Fedigan & Gay, 1979). These factors are not only of an intellectual nature; they can also be sociological, environmental, cultural, and psychological, to name only but a few. Such factors would include for example: academic aptitude, amount of previous instruction, socio-economic status, home background, parental support, school facilities, classroom climate, teacher characteristics, motivation, level of aspiration, and attitudes toward school and learning. Clearly, this list itself is non-exhaustive. But for purposes of simplicity, several of the above variables have not been included in the model as the emphasis is on the relationships among the 3 above-mentioned variables of



Note. R_a and R_b represent the residual terms for achievement and test anxiety respectively. "Home background" would include a number of indicators such as parents' occupation, family income, and parental support toward educational achievement.
"School facilities" would include indicators such as classroom climate, teacher characteristics, and access to computers.

Figure 6 **Model showing the inter-relationships among achievement, computer experience, computer anxiety, test anxiety, demographic variables and their hypothesized directional effects.**

interest in this study. As there is a multitude of causal factors affecting achievement, and it is difficult to account for the total variance of the achievement variable, the residual term, R_a, is included to indicate the effect of variables excluded in the model. Likewise, R_b represents the residual term for test anxiety. While the residuals do encompass the variables not subsumed by the model, the choice of which variables to incorporate in the model, or the decision as to which model is more tenable, depends primarily on the researcher's substantive and theoretical insights into the problem. Yet at some point, it is important to establish closure in the model and examine the relationships among a finite set of variables. Based on the correlational investigations, the results obtained in this study are consistent with the model. Future investigations using path analysis and causal modeling (Wright, 1934; Asher, 1976) can be designed to evaluate the relationships postulated by this or a modified model. In these techniques, the analysis of the data is designed to shed light on the tenability of the model formulated on the the basis of knowledge, theoretical formulations and assumptions, and logical analysis. The methods "combine the quantitative information given by the correlations with such qualitative information as may be at hand on causal relations to give a quantitative interpretation" (Wright, 1934, p. 193).

The last group of proposed studies concern the feasibility of computer-based testing. Some issues for consideration by test administrators and evaluators may relate to the planning and logistics of administration. These could be conducted to assess the practicability of this mode of test administration in the classroom and examine factors which may limit the success of implementing computer-based testing in the school classroom. For example, how do administrative logistics differ from those of paper-and-pencil administration? What kinds of problems, if any, are caused by differences in administration? One of the features of computer-based testing most appreciated by test administrators and teachers is the reduced turn-around time between completion of the test and reporting of results. Where computer

labs are available for simultaneous testing of a class, these time savings may be of practical significance in comparison to paper-and-pencil tests.

5.5 Conclusions

Despite being identical in length, item content and sequence, the computer-based test and the paper-and-pencil test were not equivalent in terms of mean achievement scores and score distribution. There was no interaction effect between mode of test administration and gender; that is, the effects of the computer-based test and paper-and-pencil test on achievement were the same for both males and females. Modifications in the way test items are presented on a computer-based test may change the strategies with which students approach the items. Extraneous variables incidental to the computer administration such as the inclination to guess on a question, the ease of getting cues from other questions, differences in test-taking flexibility, previous experience or familiarity with computers, and attitudes towards computers may change the test-taking behaviour to the extent that a student's performance on a computer-based test and a paper-and-pencil test may not be the same.

Procedural variables in the test-taking process and the manner in which the student interacts with the computer are important also in determining the affective consequences of computerized testing. If the tasks involved in taking a test on a computer are kept simple, prior computer experience has little impact on the anxiety evoked in a student while he or she is taking a test. The students were generally enthusiastic about computer-based testing. This is encouraging for the future use of computers in testing. If attitudes are favourable, it will be easier to integrate the use of computers for testing into the educational system. Nonetheless, educators must bear in mind that factors innate to computer-based testing may affect the overall test performance of a class or of certain individuals or groups of individuals. Differences in performance and attitudes towards computer-based tests and

traditional paper-and-pencil tests might also depend upon the type of test taken and the population tested.

BIBLIOGRAPHY

Allen, M.J. & Yen, W.M. (1979). Introduction to measurement theory. Monterey, CA: Brooks / Cole .

American Psychological Association (1986). Guidelines for computer-based tests and interpretation. Washington, D.C.: Author.

Apple Computer (1988). Hypercard version 1.2.2. Cupertino, CA: Author.

Asher, H.B. (1976). Causal modeling. Beverly Hills, CA: Sage Publications.

Atkinson, J.W. (1971). Motivational determinants of intellective performance and cumulative achievement. In J. W. Atkinson & J. O. Raynor: Motivation and achievement, (pp. 389-410). John Wiley & sons: Toronto.

Attisha, M. & Yazdani, M. (1984). An expert system for diagnosing children's multiplication errors. Instructional Science, 13, 79-92.

Bateson, D.J., Anderson, J.O., Dale, T., McConnell, V. & Rutherford, C. (1986). British Columbia Science Assessment 1986, General Report. Victoria, B.C. : Ministry of Education, Province of British Columbia.

Brezezinski, E.J. (1984). Microcomputers and testing: Where are we and how did we get there? Educational Measurement: Issues and Practice, 3(2), 7-10.

Brown, J.S. & Burton, R.R. (1978). Diagnostic models for procedural bugs in basic mathematical skills. Cognitive Science, 2, 155-192.

Bunderson, C.V., Inouye, D.K. & Olsen, J.B. (1989). The four generations of computerized educational measurement. In R.L. Linn (Ed.), Educational measurement, (3rd ed., pp. 367-408). New York, NY: Macmillan .

Cambre, M.A. & Cook, D.L. (1985). Computer anxiety: Definition, measurement and correlates. Journal of Educational Computing Research, 1(1), 37-54.

Chen, M. (1986). Gender and computers: The beneficial effects of experience on attitudes. Journal of Educational Computing Research, 2(3), 265-282.

Crocker, L. & Algina, J. (1986). Introduction to classical and modern test theory. New York, NY: Holt, Rinehart and Winston.

Crocker, L., Schmitt, A. & Tang, L (1988). Test anxiety and standardized achievement test performance in the middle school years. Measurement and Evaluation in Counseling and Development, 20(4), 149-57.

Denny, J.P. (1966). Effects of anxiety and intelligence in concept formation. Journal of Experimental Psychology, 72, 596-602.

Dreger, R.M. (1978). Review of the state-trait anxiety inventory. In Buros O.K. (Ed.), The Eighth mental measurements yearbook. Highland Park, NJ: The Gryphon Press.

Dusek, J.B. (1980). The development of test anxiety in children. In I.G. Sarason (Ed.), Test anxiety: Theory, research and applications (pp. 87-119). Hillsdale, NJ: Lawrence Erlbaum Associates.

Duthie, B. (1984). A critical examination of computer-administered psychological test. In Schwartz, M.D., Using computers in clinical practice: Psychotherapy and mental health applications. NY: The Haworth Press.

Eaves, R.C. & Smith, E. (1986). The effect of media and amount of microcomputer experience on examination scores. Journal of Experimental Education, 55(1), 23-26.

English, R.A., Reckase, M.D. & Patience, W.M. (1977). Application of tailored testing to achievement measurement. Behavior Research Methods and Instrumentation, 9, 158-161.

Fedigan, L. & Gay, G. (1979). School-based elements related to achievement: Elements related to student success in schooling and education, Executive summary. Alberta Minister's Advisory Committee on student achievement.

Fletcher, P. & Collins, M.A.J. (1986). Computer-administered versus written tests - advantages and disadvantages. Journal of Computers in Mathematics and Science Teaching, 6(2), 38-43.

Ghiselli, E.E., Campbell, J.P. & Zedeck, S. (1981). Measurement theory for the behavioral sciences. San Francisco, CA: W.H. Freeman & Co.

Glass, G.V. & Hopkins, K.D. (1984). Statistical methods in education and psychology. (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall, Inc.,

Glass, G.V. & Stanley, J.C. (1970). Statistical methods in education and psychology. Englewood Cliffs, NJ: Prentice-Hall.

Greaud, V.A. & Green, B.F. (1986). Equivalence of conventional and computer presentation of speed tests. Applied Psychological Measurement, 10(1), 23-34.

Green, B.F., Bock, R.D., Humphreys, L.G., Linn, R.L. & Reckase, M.D. (1984). Technical guidelines for assessing computerized adaptive tests. Journal of Educational Measurement, 21, 347-360.

Gwinn, J.F. & Beal (1988). On-line computer testing: Implementation and endorsement. Journal of Educational Technology Systems, 16(3), 239-251.

Hale, M.E., Oakey, J.R., Shaw, E.L. & Burns, J. (1985). Using computer animation in science testing. Computers in the Schools, 2(1), 83-90.

Hambleton, R.K. (1984). Using microcomputers to develop tests. Educational Measurement: Issues and Practice, 3(2), 10-14.

Hayek, L.M. & Stephens, L (1989). Factors affecting computer anxiety in high school computer science students. Journal of Computers in Mathematics and Science Teaching, 8(4), 73-76.

Hedl J.J. Jr., O'Neil, H.F. & Hansen, D.N. (1973). Affective reactions toward computer-based intelligence testing. Journal of Consulting and Clinical Psychology, 40(2), 217-222.

Hively, W., Patterson, H.L. & Page, S.H. (1968). A "universe-defined" system of arithmetic achievement tests. Journal of Educational Measurement, 5, 275-290.

Hoyt, C.J. (1941). Test reliability estimated by analysis of variance. Psychometrika, 6, 153-160.

Jacobs, R.L., Byrd, D.M. & High, W.R. (1985). Computerized testing: The Hidden Figures Test. Journal of Educational Computing Research, 1(2), 173-77.

Johanson, R.P. (1985). School computing: Some factors affecting student performance. Paper presented at the Annual meeting of the American Educational Research Association. (ERIC Document Reproduction Service No. ED 258554).

Johnson, D.F. & White, C.B. (1980). Effects of training on computerized test performance in th elderly. Journal of Applied Psychology, 65, 357-358.

Jonassen, D.H. (1986). Effects of microcomputer display on a perceptual / cognitive task. Paper presented at the annual meeting of the Association for Educational Communications and Technology. Las Vegas, NV.

Kiely, G.L., Zara, A.R., & Weiss, D.J. (1986). Equivalence of computer and paper-and-pencil Armed Services Vocational Aptitude Battery tests. (Research Report no. AFHRL-TP-86-13). Air Force Human Resources Laboratory, Brooks Air Force Base, Texas.

Lee, J.A. (1986). The effects of past computer experience on computerized aptitude test performance. Educational and Psychological Measurement, 46, 727-733.

Lee, J.A., Moreno, K.E. & Sympson, J.B. (1986). The effects of mode of test administration on test performance. Educational and Psychological Measurement, 46, 467-474.

Levin, T. & Gordon, C. (1989). Effect of gender and computer experience on attitudes toward computers. Journal of Educational Computing Research, 5(1), 69-88.

Llabre, M.M., Clements, N.E., Fitzhugh, K.B., Lancelotta, G., Mazzagatti, R.D., & Quinones, N. (1987). The effect of computer-administered testing on test anxiety and performance. Journal of Educational Computing Research, 3(4), 429-433.

Lord, F.M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum.

Loyd, B.H. & Gressard, C. (1984a). The effects of sex, age, and computer experience on computer attitudes. Association of Educational Data Systems Journal, 18, 67-77.

Loyd, B.H. & Gressard, C. (1984b). Reliability and factorial validity of computer attitude scales. Educational and Psychological Measurement, 44, 501-505.

Mandler, G. & Sarason, S.B. (1952). A study of anxiety and learning. Journal of Abnormal and Social Psychology, 47, 166-173.

Mazzeo, J. & Harvey, A.L. (1988). The equivalence of scores from automated and conventional educational and psychological tests: A review of the literature (College Board Rep. No. 88-88, ETS RR No. 88-21). Princeton, NJ. (ERIC Document Reproduction Service No. ED 304 462).

McArthur, D.L. and Choppin, B.H. (1984). Computerized diagnostic testing. Journal of Educational Measurement, 21, 391-397.

McBride, J.R. (1985). Computerized adaptive testing. Educational Leadership, 43(2), 25-28.

McKinley, R.L. & Reckase, M.D. (1984). Implementing an adaptive testing program in an instructional program environment. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Miller, L.H. (1977). A study in man-machine interaction. Proceedings of the National Computer Conference, 46, 408-421.

Millman, J. (1984). Using microcomputers to administer tests: An alternate point of view. Educational Measurement: Issues and Practice, 3(2), 20-21.

Millman, J. & Outlaw, W.S. (1978). Testing by computer. Association of Educational Data Systems Journal, 11, 57-72.

Mizokawa, D.T. & Hamlin, M.D. (1984). Guidelines for computer-managed testing, Educational Technology, 24(12), 12-17.

Moe, K.C. & Johnson, M.F. (1988). Participants' reactions to computerized testing. Journal of Educational Computing Research, 4(1), 70-86.

Nelson, L.R. (1974). Guide to LERTAP use and interpretation. Dept. of education, University of Otago, Dunedin, New Zealand.

Nelson, L.R. (1984). Using microcomputers to assess achievement and instruction. Educational Measurement: Issues and Practice, 3(2), 22-26.

Neter, J. & Wasserman, W. (1974). Applied linear statistical models. Homewood, IL: Richard D. Irwin.

Norusis, M.J. (1983). Introductory statistics guide, SPSS-X. Chicago, IL: McGraw-Hill.

Olsen, J.B., Maynes, D.D., Slawson, D. & Ho, K. (1989). Comparisons of paper-administered, computer-administered and computer adaptive achievement tests. Journal of Educational Computing Research, 5(3), 311-326.

Pedhazur, E.J. (1982). Multiple regression in behaviorial research (2nd ed.). Fort Worth, TX: Holt, Rinehart and Winston.

Plumly, L.W. Jr. & Ray, H.N. (1989). Computer administered testing in a classroom setting: An alternative. Journal of Research on Computing in Education, 22(1), 69-77.

Popovich, P.M., Hyde, K.R. & Zakrajsek, T. (1987). The development of the attitudes toward computer usage scale. Educational and Psychological Measurement, 47, 261-269.

Raub, A.C. (1981). Correlates of computer anxiety in college students. Unpublished doctoral dissertation, University of Pennsylvania.

Reckase, M.D., Carlson, J.E. & Ackerman, T.A. (1986). The effect of computer presentation on the difficulty of test items. Paper presented at the annual meeting of the American Psychological Association, Washington, D.C.

Ronau, R.N. (1986). Mathematical Diagnosis with the microcomputer: Tapping a wealth of potential, Mathematics Teacher, 79(3), 205-7.

Sachar, J.D. and J.D. Fletcher (1978). Administering paper-and-pencil tests by computer, or the medium is not always the message. In D.J. Weiss (Ed.), Proceedings of the 1977 computerized adaptive testing conference, Minneapolis, MN: University of Minnesota.

Sampson, J.P. Jr. (1983). Measurement forum. Computer-assisted testing and assessment: Current status and implications for the future, Measurement and Evaluation in Guidance, 15(3), 293-299.

Sarason, I.G. (1972). Experimental approaches to test anxiety: Attention and the uses of information. In C.D. Spielberger (Ed.), Anxiety: Current trends in theory and research, (vol.2 , pp. 383-403). New York, NY: Academic Press.

Sarason, I.G. (1978). The Test Anxiety Scale: Concept and research. In C.D. Spielberger & I.G. Sarason (Eds.), Stress and anxiety, (vol. 5), New York, NY: John Wiley and sons.

Sarason, I.G. (Ed.) (1980). Test anxiety: Theory, research & applications. Hillsdale, NJ: Lawrence Erlbaum Associates.

Sarvela, P.D. & Noonan, J.V. (1988). Testing and computer-based instruction: Psychometric implications, Educational Technology, 28(5), 17-20, May 1988.

Simonson, M.R., Maurer, M. Montag-Torardi, M. & Whitaker, M. (1987). Development of a standardized test of computer literacy and a computer anxiety index. Journal of Educational Computing Research, 3(2), 231-247.

Spielberger, C.D. (1972a). Anxiety: Current trends in theory and research, (vol. 1), New York, NY: Academic Press.

Spielberger, C.D. (1972b). Anxiety: Current trends in theory and research, (vol. 2), New York, NY: Academic Press.

Spielberger, C.D. (1977-80). Test Attitude Inventory. Palo Alto, CA: Consulting Psychologists Press.

Spielberger, C.D., Gorsuch, R.L. & Lushene, R.E. (1970). The State-trait Anxiety Inventory Test Manual. Palo Alto, CA: Consulting Psychologists Press.

Spray, J.A., Ackerman, T.A., Reckase, M.D. & Carlson, J.E. (1989). Effect of the medium of item presentation on examinee performance and item characteristics. Journal of Educational Measurement, 26(3), 261-271.

SPSS Inc. (1988). SPSS-X User's guide, (3rd ed.). Chicago, IL: Author.

Tatsuoka, K.K., Baillie, R. & Yamamoto, Y. (1982). SIGNBUG 2: An error diagnostic computer program for signed-number arithmetic on the PLATO system. Urbana-Champaign: Computer-based Education Research Laboratory, University of Illinois.

Tatsuoka, K.K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. Journal of Educational Measurement, 20(4), 345-354.

Tryon, G.S. (1980). The measurement and treatment of test anxiety. Review of Educational Research, 50(2), 343-372.

Turner, G.A. (1987). Computers in testing and assessment: Contemporary issues. International Journal of Instructional Media, 14(3), 187-197.

Ward, W.C. (1984). Using microcomputers to administer tests. Educational Measurement: Issues and Practice, 3(2), 16-20.

Ward, T.J. Jr., Hooper, S.R., Hannafin, K.M. (1989). The effect of computerized tests on the performance and attitudes of college students. Journal of Educational Computing Research, 5(3), 327-333.

Weiss, D.J. (Ed.) (1983). New horizons in testing: Latent trait theory and computerized adaptive testing. New York, NY: Academic Press.

Wilgrube, W. (1982). Computerized testing in the German Federal Armed Forces: Empirical approaches. In D.J. Weiss (Ed.), Item Response Theory and computerized adaptive testing conference proceedings, (pp.353-59). Minneapolis, MN: University of Minnesota.

Wilson, F.R., Genco, K.T. & Yager, G.G. (1985). Assessing the equivalence of paper-and-pencil vs. computerized tests: Demonstration of a promising methodology. Computers in Human Behavior, 1, 265-275.

Wine, J (1971). Test anxiety and direction of attention. Psychological Bulletin, 76, 92-104.

Wise, S.L. & Plake, B.S. (1989). Research on the effects of administering tests via computers. Educational Measurement: Issues and Practice, 8(3), 5-10.

Wise, S.L. & Wise, L.A. (1987). Comparison of computer-administered and paper-administered achievement tests with elementary school children. Computers in Human Behavior, 3, 15-20.

Wise, S.L., Barnes, L.B., Harvey, A.L. & Plake, B.S. (1989). Effects of computer anxiety and computer experience on the computer-based achievement test performance of college students. Applied Measurement in Education, 2(3), 235-241.

Wood, S. (1984). Computer use in testing and assessment. Journal of Counseling and Development, 63, 177-179.

Wright, S. (1934). The method of path coefficients. Annals of mathematical statistics, 5, 161-215.

Appendix A

INSTRUCTIONS FOR COMPUTER-BASED SCIENCE ACHIEVEMENT TEST

PAPER-AND-PENCIL SCIENCE TEST

SCIENCE ACHIEVEMENT TEST

Instructions :

1. This test consists of 34 questions. For each item, please select the most appropriate response (A, B, C, D or E) and type it into the box provided at the bottom right hand corner of the screen. To erase or change your response, press the "delete" or "backspace" key, and then type in your new answer. Press the "return" key to go on to the next question.

2. Please try as hard as you can, but if you have no idea what the answer for a question is, select the option "E. I don't know."

3. If you wish to review your responses or change them, press the left arrowkey (<---) to return to the previous question, and the right arrowkey (--->) to go on to the next question.

Press the right arrowkey (--->) to continue

Name : _____
First Last

SCIENCE ACHIEVEMENT TEST

Instructions :

1. This test consists of 34 questions. For each item, select the most appropriate response (A, B, C, D or E) and write it into the box provided at the bottom right hand corner of the question frame.

2. Indicate a response for each item. Try as hard as you can, but if you have no idea what the answer for a question is, select the option "E. I don't know".

1. Which one of the following radiations is not a part of the electromagnetic spectrum ?

A. Cosmic ray

B. X-ray

C. Gamma ray

D. Beta particle

E. I don't know

My response to this question is:

2. In guinea pigs, fur colour is dependent on only one pair of genes and black is dominant over white. If no mutations occur, what will happen if a purebred black guinea pig is crossed with a purebred white guinea pig ?

A. All of the offspring will be white.

B. All of the offspring will be black.

C. 1/2 of the offspring will be black; 1/2 will be white.

D. 3/4 of the offspring will be black; 1/4 will be white.

E. I don't know.

My response to this question is:

3. Which one of the following facts would you not use if you wished to convince a group of people that acid rain may ruin our lakes ?

A. Some chemicals form acid in the atmosphere.

B. Acid in water will damage metals.

C. Too much acid in lakes will kill creatures living there.

D. Many industries pour chemicals into the atmosphere.

E. I don't know.

My response to this question is:

4. When 2 g (grams) of zinc and 1 g of sulphur are heated together, practically no zinc or sulphur remain after the compound zinc sulphide is formed. What happens if 2 g zinc are heated with 2 g sulphur ?

A. Zinc sulphide containing approximately twice as much sulphur is formed.

B. Approximately 1 g of zinc will be left over.

C. Approximately 1 g of sulphur is left over.

D. Approximately 1 g of each is left over.

E. I don't know.

My response to this question is:

5. In the diagram below, which number represents the gall bladder ?

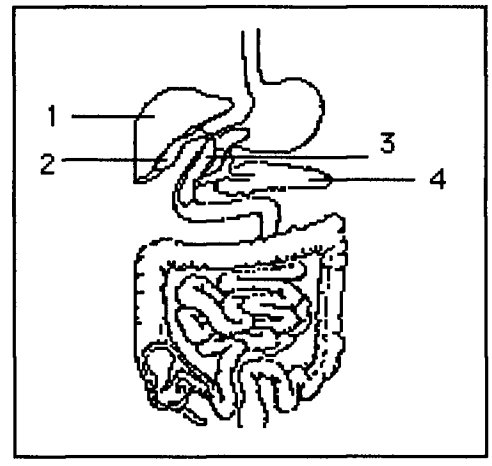
A. 1

B. 2

C. 3

D. 4

E. I don't know.



My response to this question is:

6. A pupil held a magnet to several materials and made the following table.

NAME OF	WILL ATTRACT	WILL NOT
IRON NAIL	X	
COPPER WIRE		X
SILVER COIN		X
"NICKEL COIN"	X	
STEEL SINK	X	
WOOD DOOR		X
NEWSPAPER		X
BRASS DOOR		X

What is the most logical conclusion the pupil can make about the data in the table ?

A. There is no pattern as to which materials magnets will and will not attract.
B. Magnets attract some metals.
C. Magnets attract only metals with iron in them.
D. Magnets attract metals and do not attract non-metals.
E. I don't know.

My response to this question is:

7. If you were designing a nuclear reactor and you found that it was possible to produce the required amount of energy from three different reactions, "A", "B" and "C", which reaction would you choose ?

A. Reaction "A", which produced a waste product isotope with a half-life of 20 days.

B. Reaction "B", which produced a waste product isotope with a half-life of 20 years.

C. Reaction "C", which produced a waste product isotope with a half-life of 2000 years.

D. The cheapest of the above, since half-life of the waste product is of no importance.

E. I don't know.

My response to this question is:

8. A typical Arctic food chain is:

plankton	----->	krill	----->	seals	----->	polar bears
(microscopic plants and animals)		(shrimp-like creatures)				

If thousands of polar bears were killed, what would likely happen to the krill population ?

A. Go up at first and level off.

B. Go down at first.

C. Stay the same.

D. Go up and then go down.

E. I don't know.

My response to this question is:

9. If the Earth's axis were to be tipped at an angle of 10° instead of 23°, which one of the following would be true for B.C. ?

A. The year would be longer than at present.

B. The number of June daylight hours would be fewer than at present.

C. We would be able to see both sides of the Moon.

D. The Moon would appear to be motionless rather than appearing to move in the sky.

E. I don't know.

My response to this question is:

10. The formula for the compound phosphoric acid is H_3PO_4. What is the total number of atoms in one molecule of phosphoric acid ?

A. 1

B. 3

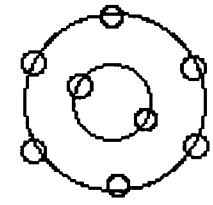
C. 7

D. 8

E. I don't know.

My response to this question is:

11. During a chemical reaction an element with 6 electrons in its outer shell, as shown here, will usually .....



A. lose electrons.

B. gain electrons.

C. neither gain nor lose electrons.

D. share electrons.

E. I don't know.

My response to this question is:

12. In plant breeding experiments, why are paper bags sometimes placed over flowers ?

A. Protect them from excessive sunlight.

B. Keep them warm.

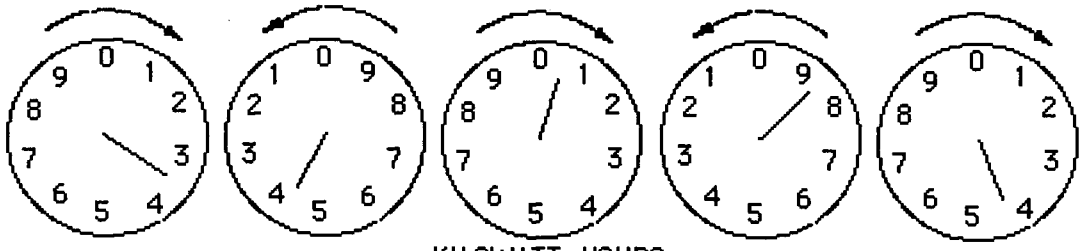
C. Prevent self-pollination.

D. Keep stray pollen away.

E. I don't know.

My response to this question is:

13.



The above meters represent the power consumption of a home on February 1. How many Kilowatt-hours are represented by the meters ?

A. 34 084

B. 35 094

C. 44 084

D. 44 094

E. I don't know.

My response to this question is:

14. In a teaching experiment fifty students were divided at random into two equal groups. One group was taught using igneous rocks; the other group used sedimentary rocks. Both groups were taught in the afternoon. In this teaching experiment, the factor which was <u>not</u> held constant was ....

A. the time of day that the groups were taught.

B. the topic of rocks.

C. the type of rocks.

D. the size of the groups.

E. I don't know.

My response to this question is:

15. A student applied force to a steel spring, and recorded the amount of stretch caused by various forces.

At which of the intervals would this graph be <u>least</u> reliable ?

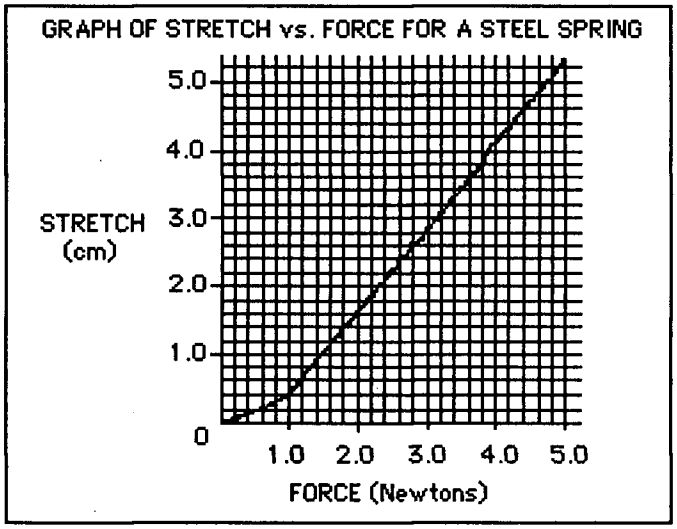
A. Between 0 and 1 N.

B. Between 1 and 2 N.

C. Between 2 and 3 N.

D. Between 3 and 4 N.

E. I don't know.



My response to this question is:

16. Using the graph below, which treatment would be <u>most</u> effective if treatments were given for only six hours ?

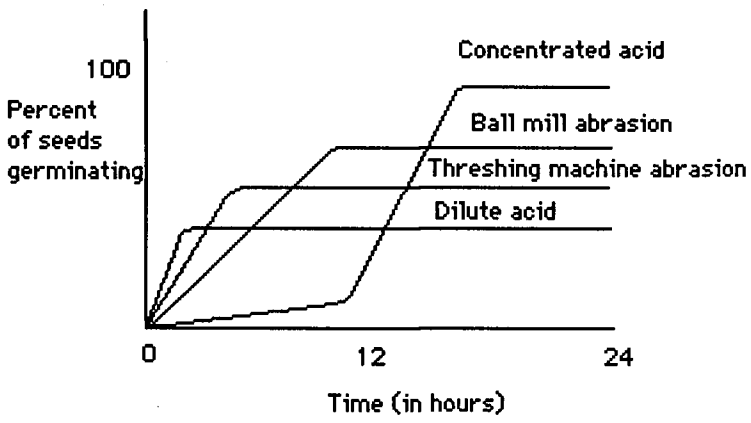
A. concentrated acid.

B. Ball mill abrasion.

C. Threshing machine abrasion.

D. dilute acid.

E. I don't know.



My response to this question is:

17. The three most important agents of erosion are .....

A. rain, rivers, and glaciers.

B. wind, soil, and ice.

C. wind, water, and glaciers.

D. water, soil, and ice.

E. I don't know.

My response to this question is:

18. Researchers have succeeded in growing new carrot plants from single mature cells of a parent carrot plant. What is <u>not</u> true about the new plants ?

A. They are clones.

B. They have exactly the same DNA as the parent plant.

C. They are a result of asexual reproduction.

D. They will produce carrots with a wide variation in colour and texture.

E. I don't know.

My response to this question is:

19. A fossil of an ocean fish was found in a rock outcrop on a mountain. This probably means that .....

A. fish once lived on the mountain.

B. the relative humidity was once very high.

C. the mountain was raised at some time after the fish died.

D. the fossil fish was probably carried to the mountain by a great flood.

E. I don't know.

My response to this question is:

20. How can excessive exposure to radiation affect future generations ?

A. Radiation could alter the genetic code in sex cells.

B. Radiation is stored in the bodies of adults and passed along to their children.

C. Radiation can be stored in eggs or sperm.

D. Radiation could alter the way your circulatory system functions.

E. I don't know.

My response to this question is:

21. Which one of the following is <u>not</u> a polluting gas ?

A. Nitrogen

B. Carbon monoxide

C. Hydrocarbons

D. Sulphur oxides

E. I don't know.

My response to this question is:

22. Two given elements combine to form a poisonous compound. Which one of the following statements about the properties of these two elements is definitely true ?

A. Neither element is poisonous.

B. At least one element is certainly poisonous.

C. One element is poisonous, the other is not.

D. Neither element need be poisonous.

E. I don't know.

My response to this question is:

23. What is the term for a substance that cannot be split into two or more substances by normal chemical means ?

A. Molecule

B. Mixture

C. Compound

D. Element

E. I don't know.

My response to this question is:

24. In which one of the following circuits will an open switch S allow bulbs Y and Z to light, but not bulb X ?

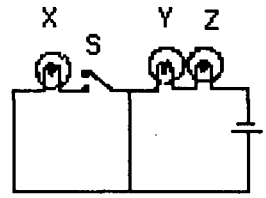
A. A

B. B

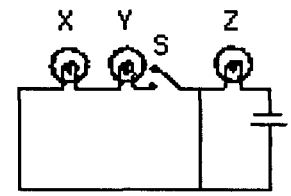
C. C

D. D

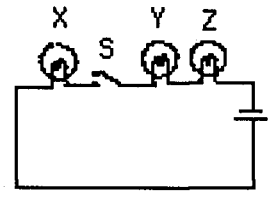
E. I don't know.



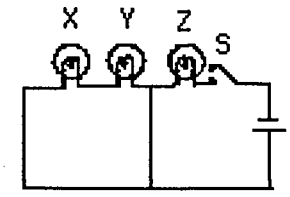
A



B



C



D

My response to this question is:

25. An electrolyte when dissolved in water breaks up into charged particles called .....

A. ions.

B. atoms.

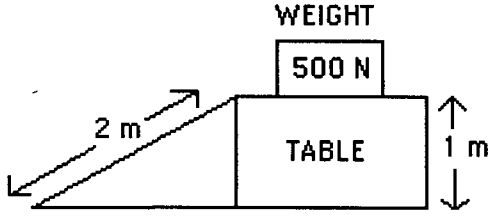
C. molecules.

D. compounds

E. I don't know.

My response to this question is:

26.



Using the inclined plane in the above diagram to move the weight from the floor to the table will .....

A. save you a small amount of work.

B. reduce the amount of effort force needed.

C. save you a lot of work.

D. double the amount of work.

E. I don't know.

My response to this question is:

27. The human embryo normally develops in the .....

A. uterus

B. abdominal cavity

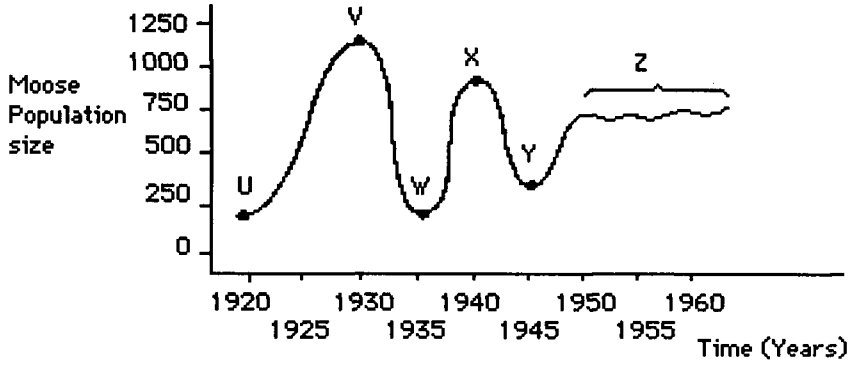
C. ovary

D. oviduct

E. I don't know.

My response to this question is:

28. This graph is a growth curve of a moose population on an island which had no natural predators of the moose.



During the time graphed where did the population's annual birth rate approximately equal its death rate ?

A. The graph does not indicate this.

B. During the time from W to X.

C. During the time from X to Y.

D. During time Z

E. I don't know.

My response to this question is:

29. What does the Milky Way consist of ?

A. A huge accumulation of dust in the solar system.

B. The trail left by a comet.

C. A cloud of gas.

D. A collection of stars.

E. I don't know.

My response to this question is:

30. As the magnification of the microscope increases, what happens to the width of the field of the image ?

A. It stays the same.

B. It increases.

C. It decreases.

D. You cannot tell what will happen.

E. I don't know.

My response to this question is:

31. Oystermen, in trying to rid their oysterbeds of starfish, dredged up many common starfish, cut each one into pieces, and dumped the pieces overboard. Oyster beds in this area soon had large populations of 'comet' starfish with one arm much longer than the others. Failure to rid the oyster beds of starfish was probably due to ...

A. the invasion of a new species of starfish.

B. the fact that "comet" starfish fed on the dead common starfish and multiplied.

C. inadequate understanding of the regenerative ability of the starfish.

D. the appearance of a new species created by mutation to fill the underpopulated by the removal of the common starfish.

E. I don't know.

My response to this question is:

32. The stars appear to rise and set because .....

A. They revolve around the earth.

B. they revolve around the sun.

C. the earth rotates on its axis.

D. the earth revolves around the sun.

E. I don't know.

My response to this question is:

33. Which one of the following parts of the cell controls cell activities ?

A. Mitochondria

B. Cytoplasm

C. Cell membrane

D. Nucleus

E. I don't know.

My response to this question is:

34. Use the table below when answering the following question.

The lowest temperature ever recorded in a coastal town was -18° C. According to the table, what percentage of antifreeze would be necessary to guarantee protection in this town ?

A. 30 %

B. 50 %

C. 90 %

D. 100 %

E. I don't know.

Freezing Point of Antifreeze / Water Solution

Percent of Antifreeze	Freezing Point in Degrees Celsius
30	- 9.5
40	- 15.4
50	- 23.0
60	- 34.7
70	- 38.9
80	- 20.8
90	- 1.6
99	+ 14.0

My response to this question is:

Appendix B

TABLE OF SPECIFICATIONS DESCRIBING THE ORGANIZATION OF SCIENCE TEST ITEMS

Table B-1 **Table of specifications describing the organization of science test items**

Domain	Topic			Total
	Physical sciences	Life sciences	Earth/space sciences	
Skills & Processes	6, 13, 15, 30 34	12, 16	14	8
Knowledge -Recall & Understand	1, 11, 23, 25	5, 21, 27, 33	17, 29	10
Application of science concepts	4, 10, 24, 26	2, 8, 18, 28	19, 32	10
Rational & critical thinking	7, 22	3, 20, 31	9	6
Total	15	13	6	34

Note. The numbers in the extreme right column and bottom row represent the number of items. The numbers in all other cells indicate the item number on the science achievement test.

Appendix C

ATTITUDE QUESTIONNAIRE

Name : _____ Sex : ☐ Male ☐ Female

First Last Date of birth : year ____ month ____

ATTITUDE QUESTIONNAIRE

The purpose of this survey is to gather information about how people generally feel about themselves, as well as their attitudes towards using computers. It should take about five minutes to complete. Please answer each statement as truthfully as possible. All responses will be confidential. Please answer every statement. Do not skip a single one. Do not spend too much time on any statement but give the answer which seems to describe how you feel.

A. SELF-EVALUATION QUESTIONNAIRE

A number of statements which people have used to describe themselves are given below. Read each statement and circle the code to the right of the statement to indicate how you generally feel. There are no right or wrong answers. Just pick the one that is really true for you.

Use these codes :

Almost never	1
Sometimes	2
Often	3
Almost always	4

For example, if your response to the following item is "Sometimes", then you would circle the number 2 like this:

When I'm bored, I like to go to a movie. 1 (2) 3 4

Use these codes : Almost never 1
Sometimes 2
Often 3
Almost always 4

1. I feel pleasant.	1	2	3	4
2. I tire quickly.	1	2	3	4
3. I feel like crying.	1	2	3	4
4. I wish I could be as happy as others seem to be.	1	2	3	4
5. I am losing out on things because I can't make up my mind soon enough.	1	2	3	4
6. I feel rested.	1	2	3	4
7. I am "calm, cool, and collected".	1	2	3	4
8. I feel that difficulties are piling up so that I cannot overcome them.	1	2	3	4
9. I worry too much over something that really doesn't matter.	1	2	3	4
10. I am happy.	1	2	3	4
11. I am inclined to take things hard.	1	2	3	4
12. I lack self-confidence.	1	2	3	4
13. I feel secure.	1	2	3	4
14. I try to avoid facing a crisis or difficulty.	1	2	3	4
15. I feel blue.	1	2	3	4
16. I am content.	1	2	3	4
17. Some unimportant thought runs through my mind and bothers me.	1	2	3	4
18. I take disappointments so keenly that I can't put them out of my mind.	1	2	3	4
19. I am a steady person.	1	2	3	4
20. I get in a state of tension or turmoil as I think over my recent concerns and interests.	1	2	3	4

B. ATTITUDES TOWARDS COMPUTERS

This section consists of a number of statements which describe how people feel towards computers. Read each statement and circle the code to the right of the statement that best describes how you feel towards computers. There are no right or wrong answers. Just pick the one that is really true for you.

Use these codes :

Strongly disagree	SD
Disagree	D
Somewhat disagree	WD
Somewhat agree	WA
Agree	A
Strongly agree	SA

1.	Computers do not scare me at all.	SD	D	WD	WA	A	SA
2.	Working with a computer would make me very nervous.	SD	D	WD	WA	A	SA
3.	I do not feel threatened when others talk about computers.	SD	D	WD	WA	A	SA
4.	I feel aggressive and hostile toward computers.	SD	D	WD	WA	A	SA
5.	It wouldn't bother me at all to take computer courses.	SD	D	WD	WA	A	SA
6.	Computers make me feel uncomfortable.	SD	D	WD	WA	A	SA
7.	I would feel at ease in a computer class.	SD	D	WD	WA	A	SA
8.	I get a sinking feeling when I think of trying to use a computer.	SD	D	WD	WA	A	SA
9.	I would feel comfortable working with a computer.	SD	D	WD	WA	A	SA
10.	Computers make me feel uneasy and confused.	SD	D	WD	WA	A	SA

*****Thank you for completing this questionnaire.*****
Please place it in the envelope provided
and return it to your teacher.

Appendix D

SURVEY OF ATTITUDES TOWARDS TESTING

Name : _____ Sex : ☐ Male ☐ Female

First Last Date of birth : year ____ month ____

SURVEY OF ATTITUDES TOWARDS TESTING

The purpose of this survey is to gather information concerning people's attitudes towards testing. It should take about two to three minutes to complete. Please answer each statement as truthfully as possible. All responses will be confidential.

This questionnaire consists of a number of statements about how you felt towards the test. Please read each statement and circle the code to the right of the statement that best describes how you felt while you were taking the test that you have just completed. There are no right or wrong answers. Just pick the one that is really true for you.

Please answer every statement. Do not skip a single one. Do not spend too much time on any statement but give the answer which seems to describe how you feel.

ATTITUDES TOWARDS TESTING

Use these codes :

Strongly disagree	SD
Disagree	D
Somewhat disagree	WD
Somewhat agree	WA
Agree	A
Strongly agree	SA

1.	I felt confident and relaxed during the test.	SD	D	WD	WA	A	SA
2.	I had an uneasy, upset feeling during the test.	SD	D	WD	WA	A	SA
3.	I froze up on the test.	SD	D	WD	WA	A	SA
4.	I was confused when working on the test.	SD	D	WD	WA	A	SA
5.	During the test, thoughts of doing poorly interfered with my concentration.	SD	D	WD	WA	A	SA
6.	I felt jittery during the test.	SD	D	WD	WA	A	SA
7.	I felt tense during the test.	SD	D	WD	WA	A	SA
8.	I wished the test did not bother me so much.	SD	D	WD	WA	A	SA
9.	I was so tense that my stomach got upset during the test.	SD	D	WD	WA	A	SA
10.	I seemed to defeat myself while working on the test.	SD	D	WD	WA	A	SA
11.	I felt panicky during the test.	SD	D	WD	WA	A	SA
12.	I worried when the test began.	SD	D	WD	WA	A	SA
13.	My heart was beating very fast during the test.	SD	D	WD	WA	A	SA
14.	I continued to worry even after the test was over.	SD	D	WD	WA	A	SA
15.	I was so nervous that I forgot facts I really knew during the test.	SD	D	WD	WA	A	SA

*****Thank you for completing this questionnaire.*****
Please place it in the envelope provided
and return it to your teacher.

Appendix E

SURVEY OF ATTITUDES TOWARDS TESTING BY COMPUTERS

Name : _____ Sex : ☐ Male ☐ Female
First Last Date of birth : year ____ month ____

SURVEY OF ATTITUDES TOWARDS TESTING BY COMPUTERS

The purpose of this survey is to gather information concerning people's attitudes towards working with and testing by computers. It should take about five to ten minutes to complete. Please answer each statement as truthfully as possible. All responses will be confidential.

Please check the boxes which apply to you.

1. Is there a computer at home that you can use ?

☐ No ☐ Yes

2. In a typical week, how many hours do you spend on using a computer ?

☐ 20 hours or more
☐ 15 - 19 hours
☐ 10 - 14 hours
☐ 5 - 9 hours
☐ 1 - 4 hours
☐ I use the computer only occasionally -- once every few weeks or so.
☐ Never -- I don't use a computer at all.
(If you checked this box, please skip question 3 and go on to the next page.)

3. During the past school year (since Sept. 1, 1989), what have you used the computer for ? (Please check all that apply)

☐ playing computerized games such as Pac-man
☐ application programs (such as in wordprocessing, database and spreadsheets)
☐ instructional programs (such as tutorial, remedial and mastery learning)
☐ computer programming
☐ courses other than computer programming courses
☐ others (please specify) _____

The next part of this questionnaire is divided into 2 sections. Section A consists of a number of statements about how you felt towards the test, and Section B consists of statements about how you felt towards testing by computers. Please read each statement and circle the code to the right of the statement that best describes how you felt while you were taking the test that you have just completed. There are no right or wrong answers. Just pick the one that is really true for you.

Please answer every statement. Do not skip a single one. Do not spend too much time on any statement but give the answer which seems to describe how you feel.

A. ATTITUDES TOWARDS TESTING

Use these codes :

Strongly disagree	SD
Disagree	D
Somewhat disagree	WD
Somewhat agree	WA
Agree	A
Strongly agree	SA

1.	I felt confident and relaxed during the test.	SD	D	WD	WA	A	SA
2.	I had an uneasy, upset feeling during the test.	SD	D	WD	WA	A	SA
3.	I froze up on the test.	SD	D	WD	WA	A	SA
4.	I was confused when working on the test.	SD	D	WD	WA	A	SA
5.	During the test, thoughts of doing poorly interfered with my concentration.	SD	D	WD	WA	A	SA
6.	I felt jittery during the test.	SD	D	WD	WA	A	SA
7.	I felt tense during the test.	SD	D	WD	WA	A	SA
8.	I wished the test did not bother me so much.	SD	D	WD	WA	A	SA
9.	I was so tense that my stomach got upset during the test.	SD	D	WD	WA	A	SA
10.	I seemed to defeat myself while working on the test.	SD	D	WD	WA	A	SA
11.	I felt panicky during the test.	SD	D	WD	WA	A	SA
12.	I worried when the test began.	SD	D	WD	WA	A	SA
13.	My heart was beating very fast during the test.	SD	D	WD	WA	A	SA
14.	I continued to worry even after the test was over.	SD	D	WD	WA	A	SA
15.	I was so nervous that I forgot facts I really knew during the test.	SD	D	WD	WA	A	SA

B. **ATTITUDES TOWARDS COMPUTER-BASED TESTING**

Use these codes :

Strongly disagree	SD
Disagree	D
Somewhat disagree	WD
Somewhat agree	WA
Agree	A
Strongly agree	SA

1. Using the computer for taking a test did not scare me at all.	SD	D	WD	WA	A	SA
2. Taking the test on a computer made me very nervous.	SD	D	WD	WA	A	SA
3. I would not feel threatened even if my classmates liked taking the test on the computer.	SD	D	WD	WA	A	SA
4. I felt aggressive and hostile toward the computer when taking the test on it.	SD	D	WD	WA	A	SA
5. I would have felt better if I had taken a paper and pencil test instead of a computerized test.	SD	D	WD	WA	A	SA
6. I felt uncomfortable using a computer for this test.	SD	D	WD	WA	A	SA
7. Now that I've finished this test by computer, I would feel at ease in taking other tests on the computer.	SD	D	WD	WA	A	SA
8. I got a sinking feeling when I saw that I had to use a computer.	SD	D	WD	WA	A	SA
9. I felt comfortable working with a computer.	SD	D	WD	WA	A	SA
10. Using a computer made me feel uneasy and confused.	SD	D	WD	WA	A	SA

1. What did you <u>like</u> most about testing by computer ?

2. What did you <u>dislike</u> most about testing by computer ?

3. Would you choose a computerized test over a paper-and-pencil test ?

☐ No ☐ Yes

Why ? _____

_____

_____

4. Do you have any other comments on how you felt about testing by computer ?

*****Thank you for completing this questionnaire.*****
Please place it in the envelope provided
and return it to your teacher.

Appendix F

CUMULATIVE PERCENTAGE DISTRIBUTIONS OF NUMBER OF EXAMINEES FOR SCIENCE TEST SCORES

Table F-1 **Cumulative percentage distributions of number of examinees for science test scores**

Score	Cumulative percent	
	Computer-based	Paper-and-pencil
0		
1		
2		
3		
4		
5		3.8
6		9.4
7	3.8	15.1
8	3.8	17.0
9	5.8	24.5
10	7.7	30.2
11	13.5	34.0
12	21.2	43.4
13	25.0	54.7
14	38.5	58.5
15	44.3	62.3
16	55.8	69.8
17	63.5	77.4
18	69.2	81.1
19	76.9	81.1
20	82.7	83.0
21	88.5	92.5
22	92.3	92.5
23	94.2	96.2
24	94.2	98.1
25	96.2	100.0
26	98.1	
27	98.1	
28	100.0	