A COMPARISON OF A COMPUTER-ADMINISTERED TEST AND A PAPER AND PENCIL TEST USING NORMALLY ACHIEVING AND MATHEMATICALLY DISABLED YOUNG CHILDREN

DISSERTATION

Presented to the Graduate Council of the
University of North Texas in Partial
Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

By

Colleen R. Swain, B.S., M.S.

Denton, Texas

May, 1997

Swain, Colleen R. A comparison of a computer-administered test and a paper and pencil test using normally achieving and mathematically disabled young children. Doctor of Philosophy (Applied Technology, Training and Development), May, 1997, 164 pp., 35 tables, 25 appendices, references, 70 titles.

This study investigated whether a computer-administered mathematics test can provide equivalent results for normal and mathematically disabled students while retaining similar psychometric characteristics of an equivalent paper and pencil version of the test. The overall purpose of the study was twofold. First, the viability of using computer administered assessment with elementary school children was examined. Second, by investigating items on the computer administered mathematics test for potential bias between normally achieving and mathematically disabled populations, it was possible to determine whether certain mathematical concepts consistently distinguish between the two ability groups. The study was conducted by administering the KeyMath-R and the CAMT to 114 third graders from private and public schools.

The results revealed no statistically significant interaction between ability group and mode of assessment between the two mathematics tests of similar content. Second, there was statistical significance in the method of assessment used, as evidenced by scores obtained on both formats of the mathematics test. Participants scored higher on all subtests of the paper and pencil format of the mathematics test than on the computer-administered format of the test. Various reasons for the difference found in mode of assessment are presented. Third, ability level was a statistically significant factor on both formats of the mathematics test. Subjects who were categorized as normally achieving in mathematics scored higher on all subtests of both tests than subjects who were categorized as mathematically disabled. Fourth, no mathematical concepts consistently distinguished between normally achieving subjects in mathematics and those who are mathematically disabled.

The study concluded that although computer-administered testing offers considerable potential, there is more to test development than changing the medium of presentation. Expressions, knowledge, ability, type of item used, and presentation all affect student performance and must be considered in test development.

A COMPARISON OF A COMPUTER-ADMINISTERED TEST AND A PAPER AND PENCIL TEST USING NORMALLY ACHIEVING AND MATHEMATICALLY DISABLED YOUNG CHILDREN

DISSERTATION

Presented to the Graduate Council of the
University of North Texas in Partial
Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

By

Colleen R. Swain, B.S., M.S.

Denton, Texas

May, 1997

TABLE OF CONTENTS

LIST OF TABLES

Table Page

CHAPTER 1

INTRODUCTION

E. L. Thorndike stated, "Whatever exists at all exists in some amount. To know it thoroughly involves knowing its quantity as well as its quality" (as cited in Crocker & Algina, 1986, p. 3). The practice of discerning what a student knows consumes a considerable amount of time in today's educational system. Measuring student ability in various constructs is paramount in modern educational settings because a major aim of education is to ascertain the child's knowledge and to use that knowledge as a cornerstone for building a personalized educational plan. Due to the vast number of students in the school systems, the goal of accurate assessment within a short amount of time is primarily carried out through the use of norm-referenced, group-administered tests. However, selecting a group-administered test over an individually administered test always involves a tradeoff. The recent infusion of computers in educational settings facilitates the application of computer technology in testing. Standardized group-administered tests, as well as individualized tests, can effectively be used with computers in the educational field.

Standardized group-administered tests, by design, accommodate a large range of ability levels. Uniformity and reduced cost are two of the advantages associated with these tests. Standard score performance distributions of these norm-referenced tests are designed to approximate the normal curve. Therefore, a majority of all examinees score within plus or minus one standard deviation of the average score, the middle 68% of the curve (Wainer, 1990). Because of this fact, the item pool for group-administered tests is weighted with items from the middle of the continuum, with only a small portion of items representing ability levels at the extremes.

The group testing structure does not easily employ the concepts of basals and ceilings (i.e., entry and exit points). Therefore, the structure of group-administered tests dictates that test examinees must complete a large number of items before reaching items that more accurately assess their ability. In such a setting, students who are achieving below normal must endure answering numerous items that are above their ability. The consequence of this situation may produce "confusion, bewilderment, and frustration to the examinee" (Wainer, 1990, p.10). This aspect of group-administered tests is opposite to what occurs in an individually administered test.

Individually administered tests are designed to limit the amount of time an examiner must spend with an individual examinee. Using basals and ceilings insures that examinees do not spend an inordinate amount of time responding to items beyond their ability level. Students enter the test at a point usually determined by previous schooling, progress either forward or backward depending upon their responses, reach the ceiling of their ability level, and exit the test.

In order to provide a more efficient manner to gain an estimate of the student's ability and still allow large numbers of examinees to be evaluated, alternative testing mode procedures have been explored. It is hoped that alternative modes of assessment would increase the precision of assessment and enable educators to meet the needs of the child in an expedient fashion while decreasing the time taken away from classroom instruction.

Modern innovations are allowing new procedures to evolve. Intertwining measurement theory with technology has resulted in advances in assessment techniques. The infusion of the computer into society has enabled testing paradigms to progress and, in some ways, to surpass expectations. "The modern performance assessment movement is based on the proposition that new testing technologies can be more direct, open ended, and educationally relevant than conventional tests, and also reliable, valid and efficient" (U.S. Congress, 1992, p. 23). This application of technology in the fields of measurement and education could alter procedures currently used in educational settings due to the numerous advantages involved in computerized testing.

Computerized assessment has many of the benefits offered by individually administered exams and group-administered exams, along with some inherent advantages over paper administered tests. First, assessment of the construct can quickly be determined. Scoring is immediate. Results from the test can be examined confidently and without the concern of a possible computational error by a test administrator (Wise & Plake, 1989). Test outcomes can be expediently turned over to the proper personnel in order for the examinee to receive treatment that caters to his or her strengths and weaknesses. Swift feedback is important to both the examinee and the test examiner. This use of applied technology may allow teachers to get a learner back on track in an expedient fashion (U.S. Congress, 1992). Second, computerized testing has been shown to require less time to provide a more precise estimate of the concept in question (Olsen, Cox, Price, Strozeski, & Vela, 1990; Wise & Plake, 1990). Third, computerized testing offers more standardization of test conditions as well as improved test security (Olsen, Maynes, Slawson, & Ho, 1989). Test files, along with student data files, can be stored securely on a networked computer system when appropriate security measures are taken. Fourth, the computerized testing paradigm offers the opportunity to gather more information about the examinee's response to an item (Olsen, 1990; Wise & Plake, 1989; Wise & Plake, 1990). For example, if desired, item response latency can be measured. Fifth, the computer potentially allows for the administration of types of questions considered impractical in the past (Wainer, 1990). Items can be presented in new and realistic ways. Because of these and other advantages of computerized testing, this paradigm is an attractive assessment option.

Computerized assessment can be appropriate for a large range of age levels (Legg & Buhr, 1992; Olsen et al., 1989). However, a majority of the research studies conducted use adults as subjects. Few studies deal with elementary age children, while even fewer studies have been performed using children with learning disabilities as subjects. By examining these two elements, computer-administered assessment and special populations, potentially great strides can be made in understanding results received from computerized assessment in the primary grades. Providing accurate and timely instructional information to the teacher will allow implementation of individually designed instructional programs for the young child with exceptional learning characteristics. This study provides a starting point for the many researchers and practitioners who are interested in the use of applied technology in the assessment of young children for the purpose of academic placement, modes of instruction, special programs, and more.

Purpose of the Study

The purpose of this study was twofold. First, the viability of using computer-administered assessment with elementary school children was examined. This study allows for a practical, real-world application of technology in educational settings. Determining the viability of computerized testing was accomplished by investigating scores received by elementary-aged students who took a mathematics test in two different formats, pencil and paper mode and computer-administered mode. The second purpose was to examine each item on the computer-administered mathematics test for bias between normally achieving and mathematically disabled populations. Salvia and Ysseldyke (1995) stated that tests must be shown to act similarly for the populations they are designed to identify.

Significance of the Study

This study has the potential to make several significant contributions to the educational field. First, the study provides additional validity evidence for computer-administered tests using young children as subjects and the results and inferences that can be drawn with this assessment tool. As computers become more ubiquitous in educational settings, this application of technology in testing has become more feasible (U.S. Congress, 1992). Second, the study demonstrates whether a computer-administered test and a paper and pencil test of similar content provide equivalent and valid psychometric information. Third, item analysis on the computer-administered test assists in determining whether certain concepts consistently distinguish between normally achieving students and students who are mathematically disabled. Fourth, the study examines whether the computer-administered test can distinguish between students who are normally achieving and those who are mathematically disabled.

Statement of the Problem

This study attempted to address the following problem: Can a computer-administered mathematics test provide equivalent results for normal and mathematically disabled students while retaining psychometric characteristics similar to the paper and pencil version?

Computerized testing provides numerous advantages to the researcher when examining diverse populations in various constructs. However, there are fewer studies using young children as subjects than there are studies with adults as subjects. Further, there is scant research available on results received from young children with learning disabilities using a computer-administered test.

Research Questions

1. Does the computer-administered mathematics test produce scores similar to those obtained by using the KeyMath-R, a pencil and paper test, with regard to overall scores and item statistics?

2. Do any test items in the item pool on the computer-administered mathematics test suggest differential item functioning (DIF) toward students of different mathematical ability level?

Assumptions

It was assumed that the computers used in the school were in an environment appropriate for student use. This included the height of the desk, the amount of room for the student to work in, correct lighting for computer use, and other ergonomic factors. It was also assumed that students had experiences with computers in educational settings.

Delimitations

The primary sample of students used in this study was from four private schools and one public school in the Dallas/Fort Worth area. Although the subjects used in the study were from similar socioeconomic backgrounds, the results cannot be generalized to students from all public and private schools in Texas or the United States. In addition, only students in the third grade were used for this study. Therefore, results could not be generalized to every grade level.

Limitations

The structure of this study did not allow for the subjects to be randomized into groups since the ability level of the subject is predetermined. Also, there was only a minimum standard for the computer hardware necessary for participation in this study. Hence, the speed of the processor, size of the monitor, and other such factors varied in the study.

In addition, this study was performed in private and public school settings. Gaining access to various schools presented a significant problem. The time of day in which the testing was conducted was at the discretion of the principal and participating teacher. Students were tested during their normal mathematics time, which was not consistent across all the schools used in the study.

Definitions of Terms

Computer-administered testing. A test administered on a computer. Computer-administered tests are not always an exact transformation of a traditional paper and pencil test to a computer presentation. Many computerized tests are able, using advanced technology, to present items in new and realistic formats yet maintain similar psychometric characteristics to comparable pencil and paper tests.

Intelligence Quotient (IQ). This study utilized students whose IQ score fell within the range of 80-120. This covers, at a minimum, 81.64% of the students in a typical classroom environment (Hinkle, Wiersma, & Jurs, 1994). These breakpoints were determined by using 1 standard deviation plus 5 points. For intelligence tests as a rule, the mean is 100 and the standard deviation is 15.

Normal mathematical achievement. A student was categorized as normally achieving when the student's score on the mathematical section of a standardized test fell within a range of 1 standard deviation plus 5 points around the expected IQ score. For example, if a student had an IQ score of 100, the range considered to be normally achieving in mathematics would be 80-120. Students who scored outside this range and were not categorized as mathematically disabled were excluded from the study.

Mathematically disabled. A student was categorized as mathematically disabled when the student's score on the mathematical section of a standardized test fell 16 points or more below the expected IQ score. By definition, a students is labeled as learning disabled if he or she is achieving more than 1 standard deviation away from the known IQ score (Texas Education Agency, 1996). Students need not be clinically labeled as mathematically disabled to fall into this category for this study.

KeyMath-Revised: A Diagnostic Inventory of Essential Mathematics (KeyMath-R). The KeyMath-R is a nationally standardized test developed by Austin J. Connolly and published by American Guidance Service. This pencil and paper test is designed for students in K-9. The test was designed to assess mathematical concepts and skills by using basals and ceilings to determine the ability level of the examinee. In this study, only four subtests of the KeyMath-R, addition, subtraction, multiplication and division, were used.

K-6 Computer-administered Mathematics Test (CAMT). A computer-administered test developed by examining the Texas Education Agency essential elements, national textbooks in mathematics, and, specifically, the KeyMath-R. The content of the CAMT is parallel to the content found in the addition, subtraction, multiplication, and division sections of the KeyMath-R. This test is designed to assess mathematical concepts and skills. This version of the CAMT contained only addition, subtraction, multiplication, and division sections.

Special population. For the purpose of this study, students who are mathematically disabled are members of the special, or exceptional, population.

CHAPTER 2

REVIEW OF RELATED LITERATURE

Introduction

Computerized testing as an assessment mode is quickly gaining popularity due to several advantages. These advantages can benefit both student and test administrator in the educational and diagnostic field. Recent developments in measurement theory and computer technology have made the computer-administered format attractive for test developers.

Regardless of the benefits available, computerized testing has not been widely used with young children or special populations. This study examines whether this testing paradigm works as well for special populations as the traditional paper and pencil testing method. This review focuses on six areas: (a) background of computerized testing, (b) advantages and disadvantages of computerized testing, (c) validity issues for computerized tests, (d) differential item bias, (e) related studies using normal populations, and (f) related studies dealing with exceptional populations.

Background of Computerized Testing

The principles of test theory, which were designed primarily as a means of examining individual differences, began in earnest in the mid 1800s (Allen & Yen, 1979). Binet and Simon "moved the study of mental testing from an academic exercise to an enterprise that could have immediate application in the classroom, clinic, and workplace" (Crocker & Algina, 1986, p. 9). The Binet Intelligence Test was the first individually administered test that would pinpoint the ability level of the examinee (Crocker & Algina, 1986). Tests which are individually administered, however, are resource intensive.

Due to the high cost associated with this individually administered testing, alternative solutions such as group testing were considered. Innovators such as Thorndyke, Thurstone, Otis, and Terman began developing new test formats, such as multiple choice (Carlson, 1994). As new developments in testing were transpiring, a pivotal event occurred in the history of mental testing. The United States entered World War I, creating a strong need for mass mental testing. Leaders in the testing field (led by Yerkes, the President of the American Psychological Association) developed the Army Alpha, which enabled the military to evaluate service men on nine mental subtests (Carlson, 1994). The benefits of this test, including increased objectivity and reduced administration costs, were quickly observed by other vocational occupations requiring mass mental testing (Wainer, 1990). Mass, or group, testing began to occur in nursing and other areas using certification tests.

Group-administered tests must accommodate a large spectrum (i.e., range) of people; hence, there must be a sufficient number of items available at every ability level. Typically, an abundance of items selected to measure the "average" examinee can be found (Olsen et al., 1989). Because the same test is given to all examinees, many items are inappropriate for a large percentage of test takers. Examinees with less ability must endure the frustrating experience of attempting to answer items that are far above their ability. An analogous situation occurs for examinees with a high level of the construct being tested. This inefficiency in the mass administration testing process is a major factor contributing to the advantages of computer-administered testing.

An extension of computer-administered testing is computer adaptive tests. In the 1900s, psychometricians began to theorize about a testing process that would improve efficiency while still maintaining standardized testing conditions (Carlson, 1994). Nevertheless, the idea of adapting a test to the individual in a large testing group had to wait until an appropriate theoretical basis was developed and the technology was available. Reckase (1989) and Carlson (1994) credited Binet for working on the first prototype adaptive test. However, in the early 1970s, Lord began working on development of a test that would preserve the standardization of mass administration tests, but with the adjustments available for individually administered tests. Lord's work was important in the field of computer adaptive testing. According to Wainer (1990), "He worked out both the theoretical structure of a mass-administered, but individually tailored test, as well as many of the practical details" (p.10). Lord played a major role in developing Item Response Theory (IRT), which is used a great deal in computer adaptive testing.

Computer adaptive testing was made possible by latent trait theories such as IRT and Rasc, and by the declining cost of computing power (Carlson, 1994). Computer adaptive testing is based on the premise that only one trait is being measured (Olsen, 1990). Assessing a single trait allows the item pool to be placed in a progressive continuum. Unidimensionality ensures that the examinee is administered items based on his or her current ability estimate and can be evaluated accurately using the least number of items possible. De Ayala (1992) remarked that unidimensionality of the item pool allows measurement errors in the estimation of an examinee's ability to be minimized.

Items calibrated with a latent trait theory such as IRT or Rasch give the examinee a 50% probability of answering the question correctly. The assumption is that examinees who possess more of the trait being measured will work harder problems than examinees with less ability. An examinee's probability of getting an item correct in no way changes the amount of a latent trait that a person possesses. "Estimation of person ability [sic] should not be affected by altering the probability of correct response" (Bergstrom, Lunz, & Gershon, 1992, p. 138). Overall, this calibration of the items using latent trait theories results in the reduction of test length.

As more computer adaptive tests have been developed and used in various professional fields, the concept of unidimensionality has come under question. New studies are being performed on whether this concept of unidimensionality is as crucial as once thought. However, many studies, such as Reckas (1989), have concluded that "unidimensionality assumptions of IRT does not necessarily require test items to measure a single ability, but rather the unidimensionality assumption requires that the test items measure the same composite of abilities" (De Ayala, 1992, p. 527).

Computerized testing continues to grow in today's educational environment for adults. The testing paradigm has become an increasingly popular method of assessment, specifically in the areas in which certification is required in order to practice the desired occupation (Lunz & Bergstrom, 1994). The popularity of computerized testing can be seen by the increased number of tests in the educational assessment marketplace. Advantages offered by computerized testing make it an attractive offer for examinees and administrators alike in various professional fields.

Advantages and Disadvantages of Computer-Administered Testing

Alternative solutions for assessment always present educators with a decision as to the most appropriate assessment tool for the school system, school, or individual child. For the educator or diagnostician to select a method different from the traditionally used one, the benefits and barriers associated with the alternative must be carefully evaluated. Computerized testing has many advantages that are appealing to both the examinee and the test administrator.

One frequently cited benefit of computer-administered testing is that the time needed to evaluate each examinee's ability is approximately half of the time required in traditional tests (Olsen et al., 1989; Wainer, 1990; Wise & Plake, 1989; Wise & Plake, 1990). This advantage is related to the point made by Wainer (1990), who noted that, with computer-administered testing, individuals can work at their own pace. Examinees can move to the next section of a test without waiting for a test administrator to start the group on the next section of the test; thus, examinees can progress as quickly or as slowly as they would like. Wainer (1990) also commented that one of the greatest benefits of computerized testing is that every individual is challenged but not discouraged.

Wise and Plake (1989) and Olsen et al. (1989) reported that computer-based testing provides the opportunity to gather more information about a testing session than merely the examinee's answers. "With computerized tests in which examinees can skip items and/or return to items on tests, information such as which items are skipped, the order in which items are answered, or how many answers are changed can be readily collected" (Wise & Plake, 1989, p. 6). The time that an examinee takes on an item (i.e., response latency) can also be tabulated and inspected. The additional information that can be gained from an examinee can assist the professional in optimizing the instructional process for the examinee.

The immediacy of scoring benefits both the examinee and those in the educational field. Examinees quickly receive feedback on their test (Olsen et al., 1989; Wise & Plake, 1989; Wise & Plake, 1990). Some tests give the student a status report with every item on the test (i.e., That is correct!). Test administrators receive a test score and a breakdown of test components when the examinee completes the exam. Since there is no test administrator per se to score the test, the concern for inaccuracy in scoring the test, along with possible subjective bias in the testing is eliminated. As documented by Groth-Marnat and Schumaker (1989), computer-based test results have the advantage of eliminating possible tester bias. Administrators, consciously or not, could assist certain groups or individuals through their mannerisms, reading of directions, or encouragement during the test. In addition, since the computer evaluates the responses given by the examinee and then generates a report, the computer is not searching for evidence to back up an educated guess.

Test security is always a consideration regardless of format. Testing material and student answer sheets must always be kept in a secured environment. Wainer (1990) concluded that test security is easier with computer-based tests than traditional paper and pencil tests. Traditional tests must be transported to and from the test site and these large boxes of testing materials are often vulnerable at various points in the transportation process. Besides the expense of shipping and handling, there is also the added concern of securing the test overnight if the test takes more than one day to complete. With computer-administered testing, test security must still be taken seriously, but many of the concerns are eliminated. Often, if a computer network is used, the test installed on the server, and the test and examinees' scores reside in one secure place. Gaining access to the test would require the correct login procedure to be performed. Once the test has been administered, gathering student scores involves downloading the files onto floppy diskettes or a secure hard drive.

Security issues are also related to the increased standardization that occurs with computerized tests (Olsen et al., 1989; Olsen et al., 1990; Wise & Plake, 1989; Wise & Plake, 1990). The computer never varies the instructions given to examinee, nor does the computing environment in which the test is given change. If a test is timed, the microcomputer can monitor the time in a precise way. The computer's internal clock is more precise than most of the watches worn by test administrators. Because of factors such as these, standardization of computerized tests is strengthened.

Cost of a computerized test can be both a benefit and a limitation. Pressman, Roche, Davey, and Firestone (1986) stated that computer-based tests can be administered by nonspecialized personnel. This reduces the cost associated with administering a computerized test. Nevertheless, development of a computerized test is an expensive venture (Wise & Plake, 1989). Wise and Plake (1990) commented that, after the initial investment for computer equipment, the cost of a computer-administered test declines if the test is repeatedly given. Lunz and Bergstrom (1994) documented that computerized testing reduces cost in relation to the printing and shipping costs associated with other printed assessment tools.

Computerized testing does have several limitations. Although these are surmountable, it is advisable to make sure that, for each situation, the benefits exceed the limitations. As previously mentioned, computer-based tests generally cost more to develop than traditional tests. With repeated use of the test, the cost diminishes, and this barrier is overcome.

In the past, the cost of the hardware needed has also been of great concern. However, as the costs of microcomputers continue to decline, this disadvantage diminishes in importance. Nevertheless, computerized testing is restricted by the availability of hardware for test administration. As computers become more common, the importance of this concern fades.

Tests that are administered by computers provide many advantages over the conventional paper and pencil test format. These advantages can prove to be extremely beneficial to individuals in the educational field. By reduction of the time needed for accurate assessment, more instructional or interaction time can be devoted to the examinee.

Validity and Computerized Tests

Educators and practitioners must be able to make appropriate inferences from the test scores on any test; hence, a meticulous explication of the validity of results should be made. According to Angoff (1988), "Validity has always been regarded as the most fundamental and important in psychometrics" (p.19). Due to the importance of this topic, it is critical to examine the validity for all tests used. This section briefly examines the concept of validity and the issues relating to computerized testing.

In Standards for Educational and Psychological Testing, published by the American Psychological Association (1985) validity is defined as "the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores. Test validation is the process of accumulating evidence to support such inferences" (p. 9). Gronlund and Linn (1990) explained that validity is the appropriateness of the interpretation of the results gathered from an instrument. Validity is not an "all or none" concept. "Consequently, we should avoid thinking of evaluation results as valid or invalid" (Gronlund & Linn, 1990, p. 49). In addition, test results should be used only for the purpose they were intended. Any other use of test results would be a gross misuse of the results. "No test is valid for all purposes" (Gronlund & Linn, 1990, p.50). As a result, practitioners, educators, and others using test results should be extremely careful with their use. The effects of psychological, educational, and ability tests are far reaching. Cronbach (1988) stated, "Because psychological and educational tests influence who gets what in society, fresh challenges follow shifts in social power of social philosophy. So validation is never finished" (p. 5).

Tangentially, computer-administered testing presents new and interesting challenges to test validation. According to Wise and Plake (1989), some unique threats to content validity arise. In the past, one threat to the validity of the computerized test was the computer hardware itself. In some situations, there was the limitation of the computer screen's displaying an entire comprehension question at once. In his study, Green (1988), proposed that, if the paragraph comprehension items are shortened to fit the computer monitor, a different construct, word knowledge, is being measured. Noijons (1994) indicated that going from one screen to another screen for a single item places additional stress on the eyes. Therefore, the content validity of those specific types of items is questioned. Green (1988) maintained that content validity is questionable only when the entire item cannot be displayed on a single screen.

Overall, providing evidence of the validity of a computerized test is much like the process for a traditional paper and pencil test. The importance of providing evidence of validity for computer-administered tests cannot be overlooked. The valuable information provided by the test developers enables the test user to determine whether the normative sample used in developing the test is appropriate for his or her population. "To use norms effectively, the tester must be sure that the norm sample is appropriate both for the purpose of testing and for the person being tested" (Salvia & Ysseldyke, 1995, p.133).

Test and Item Bias

Bias in testing is intertwined with the subject of validity. In theory, every item on a test would allow the examinee to show the depth of what is truly known about the concept being presented. However, items are often presented in such a way that examinees are prevented from demonstrating their true knowledge. When this situation occurs, the reason behind the discrepancy between recorded and true ability must be addressed. Obviously, the validity of a test would be skewed if inferences drawn for certain groups did not accurately represent that population of examinees. Brown (1983) maintained that using results with inappropriate populations biases the inferences made and contaminates assessment outcomes. When a test is referred to as biased, that means that incorrect uses of the results for certain groups lead to erroneous judgments and placements. Because of this, test bias must be a special consideration for test developers and for professionals who use tests for assessment purposes. Test developers, specifically, are obligated to attempt to show the absence of bias among various groups when the test is used correctly.

There is no universally accepted definition of bias. Further, the term is often used synonymously, though incorrectly, with differential item functioning (DIF). For clarity, item bias procedures entail both measuring effects (statistical indexes) and follow-up investigations into the sources of item difficulty (Camilli & Shepard, 1994). Because conclusions regarding bias are not purely statistical by nature, the statistical indexes have a more neutral label of DIF statistics. Simply put, DIF indexes are raw or uninterpretive relative difficulty indicators. Bias is an inference, however, that must be supported with evidence of construct validity (Camilli & Shepard, 1994). Dorans and Schmitt (1991) carefully made the distinction that "the comparison of matched or comparable groups is critical because it is important to distinguish between differences in item functioning and differences in group ability" (p. 5). It is critical that statistical findings are followed up by logical analysis to interpret the cause of differential difficulty and to determine whether or not these factors are relevant to the construct being measured.

Test bias is another term that is used in conjunction with item bias. However, one term does not necessarily imply the other. According to Camilli and Shepard (1994), test bias is “invalidity or systematic error in how a test measures for members of a particular group. Bias is systematic in the sense that it creates a distortion in test results for members of a particular group” (p. 8). This does not mean that if one group scores lower than another group the test is biased. Students who have a lower ability in a certain area would be expected to score lower than students who do not evidence that disability. The concern occurs when the test scores produce a larger distance in ability level than is known to actually exist. A test developer wants the difference in scores to be due only to the “true” difference in the ability being measured.

Test bias can occur if the format of the test items is inappropriate for certain groups. As a result of the potential for misuse, it is paramount for developers to show that their test is appropriate, not only for the typical population but also for various special populations. These populations can be based upon categorical variables such as race, gender, culture, and ability level. For example, a question that involves a cultural or regional event would not be a suitable question for examinees outside that culture or region. Most test developers attempt to avoid this by writing questions for the mainstream population (Hammill, Pearson, & Weiderholt, 1996).

There are ways for test developers to demonstrate the absences of bias in the test. As noted by Hammill et al. (1996), test developers can do the following to show absence of bias in their test. First, they can describe the content in terms of potential bias. Second, developers can include demographics on the targeted populations in the same percentages as documented in the last census. Third, by providing separate reliability and validity information for each of the various comparison groups, it is hoped that the lack of bias will be observed. Finally, test developers can show that the test items are appropriate for selected exceptional populations as well as for the mainstream population. It is this special examination of test items that absorbs much of the effort of test developers.

The selection of proper test items is an important element in the development of a test. The item pool for a test must be carefully examined for bias. As documented by Camilli and Shepard (1994), "Bias in individual test questions may be thought of as systematic error that distorts the meaning of test inferences for members of a particular group" (p. 1). Item bias procedures assist the researchers in detecting or "flagging" possible biased items. Differential item functioning (DIF*)* indices identify all items that perform differently for different groups. After logical analysis as to why the items seem to be "more difficult" for one group than for another, a subset of DIF items would be identified as biased and eliminated from the test (Camilli & Shepard, 1994; Doran & Schmitt, 1991). According to Steinberg, Thissen, and Wainer (1990), when items are different, there is evidence of multidimensionality between groups. DIF, a symptom of multidimensionality between groups, implies that something other than the attribute intended to be measured influences performance on the item (Camilli & Shepard, 1994; DeAyala, 1992). Hence, item bias detection becomes a matter of making comparisons between groups on probabilities of success on an item while controlling for ability. This detection process must be done item by item. If some of the items are biased toward certain populations, then the validity of the test given to one examinee will not match that of another examinee. Hammill et al. (1996) used Jensen's Delta Scores method and

Camilli and Shepard's three-parameter IRT approach to detect for item bias in their test. These two methods are discussed below.

DIF can use the item characteristic curve (ICC) associated with Item Response Theory (IRT). ICCs are assumed to be stable across groups and represent a strategy for detecting item bias. "The ICC represents a mathematical function that shows the probability of an examinee's getting an item correct in relation to the ability being measured" (Hammill et al., 1996, p. 56). Camilli and Shepard (1994) remarked that item bias is indicated when the ICC shows that the conditional probability of a correct response for an item is different for the two groups. ICCs for different groups can then be compared, and the areas between the equated ICCs can be used as an indicator of item bias. Camilli and Shepard (1994) documented four methods of determining the size of the potential difference between group: (a) the simple area indexes, (b) the probability differences indices, (c) the b parameter indices, and (d) the ICC method for small groups.

Jensen's Delta Scores method, which is based in classical test theory, can also be used to detect item bias. This method is often performed concurrently with the 3-parameter IRT method. Delta Scores are linear transformations of the z-scale, where Delta = 4z + 13 (Hammill et al., 1996). The Pearson product moment coefficient, r, between the Delta Scores of different groups is an indication of item bias. The fundamental assumption for this method is that, if the correlation is near 1.00, then the relative difficulty of the items was the same across the groups, and therefore the groups were measured in a similar fashion. Correlation coefficients that are significantly less than 1.00 are considered to be an indication of DIF (Camilli & Shepard, 1994).

Finally, item bias and test bias are matters of considerable importance in the development and use of a test and item bias procedures can and should be used from the beginning of the test-development process. A significant DIF index does not automatically denote bias; only if the item is assessing a factor irrelevant to the concept being measured would the item be labeled as bias. Statistical tests can be performed and analyzed to determine which items should be included in the item pool. At the conclusion of the development process, test bias must be addressed. A critical issue that is often overlooked, according to Hammill et al. (1996), is to ensure that test results stand, not just for normal populations but also for exceptional populations. Thus, it is prudent for test developers to spend considerable time providing evidence that their tests have an absence of bias for diverse populations.

Computerized Test Studies With Normal Populations

Computerized tests using normal populations have provided researchers with much information. Numerous studies have accumulated evidence for the validity of computerized tests as alternatives to the traditional paper and pencil tests. Computerized tests have been developed and validated in many psychological and academic areas. Some examples of how these tests have been used are for vocational aptitude tests (Moreno, Wetzel, McBride, & Weiss, 1984), a complete test battery (Henly, Klebe, McBride, & Cudeck, 1989), college placement tests (Hsu & Shermis, 1989; Stocking, 1987), nursing certification tests (Lunz, Bergstrom, & Wright, 1992), and music listening tests (Vispoel & Coffman, 1992). These and other studies have enabled researchers to increase the computerized testing knowledge base and attempt to use this evaluation format in new areas. The reasons for test development in numerous fields are varied.

Kumar and Helgeson (1995) reported that calls for reform in science education and the efforts to improve assessment techniques have contributed to the increase of computerized tests in the field. "There is a growing emphasis on a systematic research and development of alternative assessment methods to evaluate the processes, instead of the products, of learning science" (Kumar & Helgeson, 1995, p. 29).

Computerized assessment has also infiltrated the reading field. However, Wepner (1991) noted that there have been small numbers of computerized assessment packages in reading. She stated that "this slow growth stems partially from our profession's quandaries about what should be assessed and how" (p. 62).

Meta analysis has been performed on the literature (Bergstrom, 1992; Bugbee & Bernt, 1990) dealing with computer-administered tests using adults as subjects. For example, in the report by Bugbee and Bernt (1990), 6 years of findings from the use of computer-administered testing at the American College were examined. Results indicated that student performance on computer-administered exams were as good as that found on paper and pencil exams. It was also noted that students consistently gave computer-administered testing a positive rating. Results showing no significant difference between computer-administered tests and paper and pencil tests are supported by the findings in various other studies (Dimcock & Cormier, 1991; Fletcher & Collins, 1986-1987).

Regardless of the fact that studies using computerized tests have not developed equally in various fields, the knowledge that can be gained from these studies is considerable. Described below are computer-administered and computer adaptive studies found in the literature using young children who are categorized as normal as the research subjects.

McDonald, Beal, and Ayers (1987, 1992) conducted a study on whole number computational skills, which was first reported in 1987 and then again in 1992. This research project examined the performance of subjects in Grades 3 through 6 on a traditional test and a computer-administered test. The authors concluded that by comparing performance between the two modes of assessment, it would be possible to determine if the computer-administered test provided a similar profile of the student's ability.

Students used in the study were randomly assigned to treatment order so that all students would have a turn on the computer. Items used on both formats of the tests were from the Diagnostic Test of Computational Skills. Items were tailored to the student's grade level. Third graders were tested on addition and subtraction of whole numbers; fourth graders were tested on subtraction and multiplication of whole numbers; and fifth and sixth graders were tested on multiplication and division of whole numbers. Students on both tests were unable to review previous items but were allowed to use scratch paper to complete calculations. Students took the tests within a 1 week period.

Results from this study showed that computer performance was almost identical to that of the paper and pencil performance. Additional analysis of the results showed that students made almost exactly the same number and type of errors on the two tests. In addition, it was noted that the more text on the screen the slower the performance. Finally, the more complex the keyboard entry, the slower the performance was.

Researchers concluded that "the computer can become a valuable ally for the teacher and diagnostician" (McDonald et al., 1992, p. 26).

Olsen, a leader in computerized testing with young children, also performed studies using young children as subjects. In one California study conducted during 1989, a comparison of three different modes of test administrations was made. These modes were paper administered tests, computer-administered, and computer adaptive tests. Tests used in the study were prepared from the mathematics section of the California Assessment Program item bank. These tests were in a multiple-choice format. Items were selected for third graders (N=350) and sixth graders (N=225). There were 55 items at Grade 3 and 62 items at Grade 6. Students were randomly assigned to one of four experimental groups. Each group received two different testing formats according to a specified design, as shown in Table 1.

Table 1

Group Assignment for Testing Format

Group	First test	Second test
1	Computer-administered	Computer-adaptive
2	Computer-adaptive	Computer-administered
3	Paper-administered	Computer-adaptive
4	Computer-adaptive	Paper-administered

The paper-administered and computer-administered tests were scored with a "number right" rubric, whereas the computer-adaptive test was scored with the ability (theta) rubric. Therefore, all item and test scores were converted to the same common ability scale. Once this conversion occurred, a repeated measures analysis of variance was calculated for paper-administered and computer-administered tests. "The repeated measure analysis adjusts for individual student variation by examining deviations from the individual's overall average test score performance" (Olsen et al., 1989, p. 316). For the third-grade group, there were no statistically significant results for test mode, order, or test mode X order interaction effects. For the sixth-grade group, there was only a statistically significant order effect found. Ability estimates using IRT were also calculated for all three test administrations. According to Olsen et al. (1989), "These data show very similar and approximately normal distributions for each method of test administration. The computer-administered and computer adaptive tests show slightly higher means and slightly lower standard deviations than the paper-administered test" (p. 319). It was stated that this pattern was found for both third and sixth graders. A Pearson product moment correlation coefficient of .87 was found between the paper and pencil test and the computer-administered test.

Olsen et al. (1989) concluded that the results of this study gave educators many issues to consider. Results from the study certainly support the use of computer-administered tests in areas with young children.

> With the increasing number of microcomputers which are being purchased by schools and districts nationwide, this study envisions that operational computerized testing systems can be developed and implemented for schools and district used in administering instructional teacher-made tests, achievement tests and district and

> statewide assessment testing. (Olsen et al., 1989, p.323).

Olsen (1990) performed another study in Utah with 72 sixth and seventh grade students, ages 11 and 12. This study compared the results from a computer-adaptive version of the Weschler Intelligence Scale for Children-Revised (WISC-R) with the results from the individually administered form of the WISC-R. Analyzed results suggested a highly significant correlation between the two tests. "This means that such computerized achievement or assessment tests can be given more frequently without any significant loss of allocated instructional time" (Olsen, 1990, p. 36).

Bronson (1985) also performed a study designed to determine whether the use of technology was even a viable option with the very young child. The study used children, ages 3 to 5, and observed the differences in reaction to a test administered by an adult and a test administered by a computer. The children responded to the technology

> much as they responded to the human testers except that: 1) they obviously enjoyed the [computer] situation much more, 2) they appeared to be more interested in the various items and more able to concentrate, and 3) they seemed less stressed when they made errors and more able to continue cheerfully and with high motivation (p. 10).

This again supports the assumption that it is acceptable to use technology to assess young children.

These studies strongly indicate that the computer-administered test is an appropriate mode of assessment for young children in the mainstream (normal) population. Through use of a computer-administered test to assess young children's understanding of specific subjects, appropriate adjustments can be made to the curriculum in order to help the child in an expedient fashion.

The assumption cannot be made that the traditional paper and pencil test and the computerized test will produce comparable results for all groups of subjects. Therefore, studies involving subpopulations must be investigated to determine whether this pattern remains constant.

Computerized Test Studies With Exceptional Populations

Responses of special populations cannot be assumed to be the same or similar to responses given by a normal population. Therefore, in order for researchers, diagnosticians, and teachers to have confidence in the inferences drawn from assessment tools, studies must be performed with special populations to determine whether the results from computerized tests are comparable to those produced on traditional paper and pencil tests.

Research studies involving exceptional populations with computerized tests produce varied results. Studies involving learning disabled students, at-risk students, patients with brain damage, and patients from psychiatric institutions are examined. The potential that these and other computerized tests offer special populations has captured the attention of those working with these populations. For example, Fuchs, Fuchs, and Hamlett (1992) stated that "teachers can use the tools to enhance their instructional decisions and their student's achievement" (p. 60).

Studies using adult subjects still dominate the literature. One such research study by Engdahl (1991) was conducted in Minnesota. The overall focus was to examine three administrative conditions of the Differential Aptitude Test (DAT). The administrative modes were paper and pencil, fixed-length computer adaptive, and variable-length computer adaptive tests. Subjects were classified into the following categories: medical, mentally ill, chemically dependent, brain injury, and no disability.

Results from Engdahl's (1991) study showed that subjects scored higher on the computer-adaptive tests than on the paper and pencil test. Subject satisfaction was also higher for the computerized versions of the DAT. Several interesting points were made by the researchers in this study. First, the researchers were unable statistically and logically to account for the difference in scores between the paper and computerized test. They questioned whether the "initial equating process" was appropriate for this unusual population. Second, results indicated that the computerized versions of the DAT produced limited time savings. This was believed to be the case because of the population. Generally speaking, the more severe the disability, the more time the adult subjects took on each test item.

When considering special populations, it is important that the assessment tool be able to distinguish between the populations. Allen, Ellinwood, and Logue (1993) also used adults as subjects in their study, which compared the use of a computer-assisted battery of neuropsychological tests (CNT) with psychiatric inpatients and normal volunteers. Researchers concluded that the CNT battery did discriminate cognitively impaired patients from nonimpaired patients.

A final study using adults as subjects was conducted by Berger, Chibnall, and Gfeller (1994). This study compared the standard and computerized versions of the

Halstead Category Test, which is sensitive to brain dysfunction. Subjects included 95 adult patients, mean age of 29.7, in a private psychological rehabilitation clinic. Results from this study showed that subjects who completed the computer version of the test made significantly more errors than those taking the standard version. It was noted that subjects were not randomly assigned to a treatment, so the group taking the computer version could have been significantly different from the group taking the standard version.

These studies instill some confidence in using computerized tests with special populations. Results from the studies show that different factors must be considered to determine whether computerized testing is effective for the special population of interest. However, the combination of age and exceptional population must be examined. Studies have been conducted that compare using computer-administered tests and the traditional paper and pencil tests with both gifted children and children with behavioral problems (Katz & Dalby, 1981) and with physically disabled persons (Wilson, Thompson, & Wylie, 1982). The following studies include young children and youth as subjects.

Pressman et al. (1986) conducted a study with 40 learning disabled boys, ages 7 through 11, and a control group of boys of the same age without learning disabilities. The purpose of the study was to determine whether the computer-administered test could distinguish between subjects with and without learning disabilities The learning disabled subjects were selected "on the basis of a primary diagnosis of learning disability resulting from comprehensive medical and psychological screening and assessment procedures" (Pressman et al., 1986, p. 485). The study examined scores from the computerized Goldman-Fristoe-Woodcock Auditory Skills Test battery. It was shown that the test was able to successfully distinguish between the two groups of subjects. Therefore, evidence indicates that computer-administered tests should be able to be used with nonnormal populations.

Signer (1991) orchestrated a study using a computer-administered mathematics tests with high school students who were labeled at-risk. These students had a history of high absenteeism and academic setbacks. As students progressed through the at-risk academic program, they were continually tested by taking the computerized mathematics test. At the conclusion of the study, an increase in mathematical achievement was noted by the students. Findings such as increased motivation, self-confidence, and self-discipline supported results that were similar to those found with subjects from a normal population.

Although these studies do not overwhelmingly support the use of computerized tests with special populations, they provide a support base for researchers of today and the future. As more computerized tests are developed and normed with exceptional populations, the use of computerized tests with these populations could grow at exponential rates.

Conclusions

Computer-administered testing provide many advantages for the examinees, test administrators, and other professionals in the educational and diagnostic fields. However, it has been discovered that there are few studies involving computerized testing using young exceptional populations. As noted by Hammill et al. (1996), most authors attempt to write items for the "mainstream" population. This is problematic for test users of exceptional populations because it casts the validity of the test results into doubt. Tests used for special populations should have norms that are stable across normal and nonnormal populations.

> Scores based on the performance of unrepresentative norms lead to incorrect estimates of relative standing in the general population. To the extent that the normative sample is systematically unrepresentative, in either central tendency or variability, the inferences based on such scores are incorrect and invalid. (Salvia & Ysseldyke, 1995, p. 173)

Because of this gap in the literature in the area of computerized testing, this study has examined an application of current technology, a computer-administered test using subjects with a learning disability. Besides the need for more studies involving special populations, this review of literature revealed several additional areas of concern. First, additional research is needed on computer-administered tests with young children. Although special populations are found in this review of literature, few involve young children with learning disabilities. Second, there is a demand for more studies to determine whether the use of computer technology, specifically computerized testing, can distinguish between normally achieving subjects and subjects with learning disabilities. This study attempts to address the areas of concern that are apparent in this review of literature.

CHAPTER 3

METHODOLOGY

Purpose of the Study

This study was designed to examine the viability of using a computer-administered mathematics test with elementary aged children from both normal and learning disabled populations. Scores received by third graders on a pencil and paper mathematics test and a computer-administered mathematics test were investigated, as well as an analysis of potential item bias on the computer-administered test. The examination of items according to ability level assisted in exploring whether there were some concepts that consistently distinguish between students who are normally achieving in mathematics and those who are mathematically disabled.

Measured variables of interest in the study included (a) results received on the four sections of the mathematics test using both testing formats; and (b) differences in items by ability group on the computer-administered scores as measured by item difficulty, item discrimination, either a 2-parameter item characteristic curve (ICC) or 3-parameter ICC technique, and a Delta plot analysis. The three categorical independent variables were learning ability, mode of assessment, and mathematical subtest.

Population

The population for this study was third-grade students in both private and public schools in the North Texas area who are either normally achieving in mathematics or mathematically disabled. This population included students who were between the ages of 7 and 10. Participating in the study were 4 private schools in the Dallas area and 1 public school in the Mid-Cities area of Dallas/Fort Worth. Three of the private schools maintain an admissions policy stating that students attending the school must have diagnosed learning disabilities. However, this does not mean that all research participants from these schools were mathematically disabled. The remaining private school was a parochial school that admitted students with various ability levels. The public school used in the study serviced students who lived in privately owned homes and apartments in the surrounding area.

Sample

For the first research question, a total of 114 subjects in the third grade participated in this study. The average age of the subjects was 9 years, 1 month old. The youngest subject was 7 years, 8 months old, and the oldest subject was 10 years, 10 months old.

By gathering subjects' zip codes, it was possible to access the 1990 national census data to obtain an overall demographic picture of the sample used in the study. Information on race, family income, language, and education level are reported in Tables 2 through 5. This allowed the researcher to conclude that a majority of the sample are white, middle-to-upper income, English-speaking students where education is perceived as important.

Table 2

Sample by Race

Race	Mean (%)	Standard deviation
White	87.37	10.05
Black	5.35	5.94
Indian	0.34	0.14
Asian	3.52	2.06
Other	3.41	4.57

Table 3

Family Income for Sample

Income	Mean (%)	Standard deviation
<15,239	7.20	3.55
15,240 - 25,300	10.08	4.12
25,301 - 35,499	14.76	3.75
35,500 - 50,749	20.16	5.63
50,750 - 74,999	27.34	6.98
>75,000	20.45	10.63

Table 4

Sample by Language

Language	Mean (%)	Standard deviation
English Only	87.63	6.79

(table continues)

Language	Mean (%)	Standard deviation
Other Language	12.36	6.79
English is poor	16.45	9.04

Table 5

Sample by Education

Education Level	Mean (%)	Standard deviation
High School	39.92	15.92
Bachelor's degree	45.96	11.48
Graduate degree	14.11	4.74

The control group consisted of 65 subjects categorized as normally achieving in mathematics. The experimental group contained 49 subjects classified as mathematically disabled.

The second research question required a large control group (N > 100) because of the statistical test and software package being used (BIMAIN, 1994). Since the original sample contained only 65 subjects in the control group, this group was doubled. The experimental group (N = 49) remained the same. This results in the sample's increasing to N = 170 and makes the sample contrived. This doubling was considered justifiable because the research question deals with only statistical differences between two ability groups on one of the methods of assessment. The technique of contrived or simulated data is prevalent in studies dealing with new statistical methods, theories on computer adaptive testing, and Monte Carlo situations (Feinstein, 1995; Law, 1995).

Subject Categorization

Participants in the study were placed in one of two ability-level categories: normally achieving in mathematics or mathematically disabled. Subjects were categorized by examining the intelligence quotient (IQ) and the arithmetic score from the standardized mathematics exam used by the school for assessment purposes. These items were obtained by the school principal or teacher from the subject's permanent record. By examining these two scores, the investigator was able to determine whether the research participant was performing at the expected level for his or her ability level. Subjects excluded from this study were those whose mathematics score deviated by more than plus or minus two standard deviations from their expected score. Subjects were classified as mathematically disabled if their scores indicated a performance of more than one standard deviation below the expected score. There were 49 subjects in the study classified as mathematically disabled. All other subjects, except those excluded from the study, were categorized as normally achieving in mathematics. There were 65 subjects classified as normally achieving in this study.

General Design

This quasiexperimental study utilized a method that most resembles the nonequivalent control group design as described by Campbell and Stanley (1963). Students who were normally achieving in mathematics served as the control group in this study. As reported in the review of literature, research suggests that no difference would be expected between the test results from the paper-administered test and that of the computer-administered test. This conclusion, however, has been drawn primarily with subjects who are considered to be part of a normal population. This study included an exceptional population, students who are mathematically disabled, which served as the experimental group for the study. Research participants took an addition, subtraction, multiplication, and division component in each test format (paper and pencil versus computer-administered).

Sample Size

According to Cohen (1988), the sample size for this experiment would be 20 per group. This estimate was achieved by having an alpha level of .05, an estimated medium effect size of .25, and a power level of .80. This would result in N=40 for the entire study. Kraemer and Thiemann (1987) remarked that for a two-tailed test at the .01 level, the sample size for each group would need to be 35, resulting in N=70. In order to avoid finding results that border on being found by chance, the initial desired sample size for each group is 40, which would have resulted in an N=80.

The sample that was used for this study had a total of 114 subjects, with 65 subjects in the normally achieving group and the remaining 49 subjects in the group of mathematically disabled subjects. This sample exceeded the sample size recommended by both Cohen (1988) and Kraemer and Thiemann (1987); therefore, there can be some confidence in the power and effect size found in the study without concern for the problems associated with a small sample size.

Instrumentation

The measures used in any study significantly contribute to the strength of the overall study. Demonstrating the rigor of the instrument includes discussing the reliability and validity of the measure, the intended population, and the purpose for which the instrument was designed. The two measures used in this study were the KeyMath-R test and the K-6 Computer-administered Mathematics Test.

KeyMath-R

KeyMath Revised: A Diagnostic Inventory of Essential Mathematics (KeyMath-R) is a test that evaluates a student's understanding of fundamental mathematical concepts and skills. KeyMath-R was developed by Connolly (1988) and published by the American Guidance Service. The 258-item instrument is designed to evaluate a child's knowledge across 13 strands in three general areas: (a) basic concepts, (b) operations, and (c) applications. This test is designed to be used with students in grade levels K-9. For the purpose of this study, only four sections of the KeyMath-R test-- addition, subtraction, multiplication and division-- were used. Test items for these sections of the test can be found in Appendix A.

It is important to know the reliability of the instrument being used. Reliability deals with the consistency of the scores obtained from using the instrument. The KeyMath-R manual provided researchers with evidence of reliability using three methods: alternate-form, split-half, and an IRT method. For alternate-form reliability, the total test has a .90 reliability coefficient. Internal consistency for the instrument was obtained by correlating the odd and even problems. Reliability coefficients for the subtests range from .70-.80. Total test coefficients were from the mid- to high .90s for the age categories. The Rasch model, a form of IRT, produced subtests coefficients of .70-.80 and total scores of mid- to high .90s. Sattler (1992) stated that reliability coefficients of .80 and higher are generally considered to be acceptable for standardized tests. According to this standard, scores from the KeyMath-R are reliable when used correctly.

Although the reliability of a test is important, the validity of a test is critical. Validity determines whether the test measures what it purports to measure. Without validity, administering a test is pointless.

For evidence of the content validity in the KeyMath-R, subject matter experts worked on three main objectives.

> 1. The creation of a comprehensive blueprint reflecting essential mathematics content, existing curricular priorities, and national trends.
> 2. The specification of that blueprint into carefully described content segments (domains) of relatively equal values, upon which to judge student mastery.
> 3. The development of sets of items that accurately assess student mastery of the specified content. (Connolly, 1988, p.72)

Construct validity was determined in two ways. Since the test is to measure mathematically ability, it is natural to assume that, as students develop, their mathematical ability will develop as well. Therefore, developmental change was examined by giving Form A in the fall and Form B in the spring. Internal consistency was demonstrated showing intercorrelations among subtests. Once again, these findings demonstrated that the test scores for the KeyMath-R test were valid. Connolly (1988) reported that test scores were also compared and found to be consistent with those of the Comprehensive

Test of Basic Skills (CTBS) and the Iowa Tests of Basic Skills (ITBS).

The KeyMath-R test can be used for five main reasons: assessment for general instruction, assessment for remedial instruction, contribution to global assessment, pre- and posttesting, and curriculum assessment.

A great deal of normative information is provided by the manual. This includes the range of scores, the means, and the standard deviations. Standard error is also furnished. Administrators of the KeyMath-R can select an alpha level with which they feel comfortable and can determine the appropriate confidence interval. The types of scales provided are standard scores, percentile ranks, stanines, normal curve equivalents, grade equivalents, and age equivalents.

The sample of students used in the development of this test came from the testing of children in 19 states. Census reports were used to approximate the K-9 school age population. The samples of children were proportionate to those found in the census across geographic regions of the United States. In each geographic region, the sample was proportionately divided by grade, gender, socioeconomic level, and race. No mention of a specific exceptional population was found in the KeyMath-R administrator material. However, there was mention of the inclusion of special populations in the overall sample. Nevertheless, this test was selected due to its validity and reliability and the instrument's heavy use in special education and with special populations in recent school years.

K-6 Computer-administered Mathematics Test

The K-6 Computer-administered Mathematics Test (CAMT) is a computer-administered test that was designed to assess a student's ability in addition, subtraction, multiplication, and division. This test is currently unpublished. The CAMT contains four sections: addition, subtraction, multiplication, and division. Evidence of reliability and validity are continually being accumulated and are further discussed below.

The item pool for the CAMT was created by a thorough study of the essential elements for mathematics as established by the Texas Education Agency. In addition, textbooks by elementary level publishers approved by the Texas Education Agency were carefully analyzed. In order for this test to be studied along with the KeyMath-R, the conceptual algorithm of each item in the addition section of the KeyMath-R was matched with an item in the CAMT. Items on the CAMT can be examined in Appendix B.

Face validity and content validity were established by having experienced mathematics teachers in the elementary schools examine the test content as well as how the items appear. Suggestions made by the teachers , such as decimal alignment, item selections, and response formats, were included in the CAMT. Additional evidence of content validity has been address with the examination of item difficulty and item discrimination. This information can be found in Appendixes D, H, L, and P. Criterion validity was addressed by correlating the scores obtained on the test with the KeyMath-R, an existing mathematics test. The correlation coefficient between the two tests is .7289.

The CAMT was designed to run on a Windows® based 386 or above computer and on a Macintosh® computer. Careful design considerations were made throughout the development process so that the CAMT would be equivalent on the two different computer platforms. The interface of the test was designed according to standard human computer interface (HCI) specifications, as recommended by textbooks, publishers, and leaders in the field of HCI. Guidelines set forth by Shneiderman (1992), a leader in the field of HCI, were strong factors in how the interface looks in the CAMT. Examinees type in their answer for each item, and once an examinee is satisfied by the response, he or she clicks the OK button to proceed to the next item. Examinees use a combination of the keyboard and the mouse to activate the program. The instructions and activation buttons are in a consistent location on the screen, regardless of the type of item seen by the examinee. For the Windows® version of the CAMT, the built-in computer calculator is disabled by the program. The schools using Macintosh® computers had a security system that prohibited students from switching between different programs.

The Windows® version of the CAMT was generated in the Visual Basic programming language. The program is set to display the item in the maximum space allowed on the monitor. The font used is MS San Serif, which is packaged with Windows®. This ensures that all subjects see items in a similar way. The only difference in viewing the items is the size of the monitor. The Macintosh® version was created in HyperCard authoring language. The font used was Helvetica, which is packaged with most Macintosh® computers. Sample screens from the CAMT can be seen in Appendix C.

Equivalence of Instruments

The content of the CAMT that was used in this study directly corresponds to the test items found in the KeyMath-R test. Both tests contain an addition, subtraction, multiplication, and division segment. Since only third graders were tested, many of the items in the KeyMath-R and CAMT were not necessary due to mathematical ability. In order for the two forms to be as equivalent as possible, problems from the KeyMath-R were used as a selection guideline. For every problem in the KeyMath-R, an item using the same mathematical algorithm was selected from the CAMT item pool to insure content equivalence.

In order to show the equivalence of the content of the two instruments, several experienced mathematics teachers examined the two tests side by side to insure that each item assessed the same concept. All agreed that the content being assessed was equivalent. The overall correlation coefficient between the two test formats is .7289. Correlation coefficients between the two different formats (paper and pencil vs. computer) and the four subtests were also found. The correlation coefficients on the subtests range from .6508 to .8221. In addition, correlation coefficients on the four subtests for each ability group were calculated. For normally achieving subjects, the correlation coefficients range from .5047 to .7935, and scores from mathematically disabled subjects produced correlation coefficients ranging from .5329 to .7654. Information for each of the four subtests appears in Table 6.

Data Collection

After the researcher received the required consent forms, the two tests were administered to each subject. Once students at a site were selected for inclusion in the study, a number, beginning with 101, was assigned to the subject. The teachers provided the investigator with the subject's birthday, IQ score, the most recent standardized score on a mathematics exam, and zip code. Teachers obtained the desired information so that the investigator would never know the identity of the subject.

Table 6

Correlation Coefficients for Instruments

KeyMath-R		Addition	Subtraction	Multiplication	Division
CAMT					
Addition	A	.6508			
	NA	.5047			
	MD	.7310			
Subtraction	A		.8221		
	NA		.7935		
	MD		.7654		
Multiplication	A			.7725	
	NA			.7218	
	MD			.7308	
Division	A				.6505
	NA				.6527
	MD				.5329

Note. A = All subjects; NA = Normally achieving; MD = Mathematically disabled.

The order of the treatments was determined through the coordination of the principal and the teacher and the availability of the computer lab. Students at the public school and two of the private schools were administered the paper and pencil test first and the computer-administered test second. At the remaining two private schools, the computer test was administered first and the paper and pencil test second. The varied ordering of the treatments allowed the effect produced from order of the treatments to be reduced.

Administration of the second test varied from school to school due to scheduling requirements of the computer lab. The shortest time between testing sessions was 1 day and the longest time between testings was 4 days. The short time between testing formats was requested by the investigator in order to reduce the possible confounding effect of intellectual maturation and testing effect on the part of the subject (Ferguson, 1981).

The KeyMath-R is a paper and pencil test that was administered by the teacher during the subjects' normal mathematics class time. All subjects received at least 1 hour to complete this test; no research participant was forced to stop working on a test due to time restraints.

The CAMT is administered on a computer. All responses given by the examinee are written to an ASCII data file. For this study, there was a facilitator in the room who started the examinee on the test, answered any questions the examinee had while taking the test, and monitored the overall administration of the test. At two of the private schools, the facilitator was the subjects classroom teacher. However, at the public school and two of the private schools, the teachers and principal requested that the researcher be present when the computer test was administered.

Statistical Analysis

The two research questions in this study required that several statistical procedures be performed. As a result, this section is a discussion of the various procedures needed to thoroughly answer the research question. Statistical analysis was performed using SAS, SPSS, and BIMAIN statistical software packages.

Upon completion of data collection and coding, the reasonableness of the data was checked. Descriptive statistics were performed to check for data-recording errors, anomalies in the data, and assumptions of the various statistical tests used in the data analysis.

The first research question examined was whether the computer-administered test and the traditional paper-administered test produced similar results according to ability level. The statistical design to be used for this process was the two-factor, fixed effects analysis of variance. A two-way analysis of variance was performed for each of the four mathematical algorithms (addition, subtraction, multiplication, division). The assumptions for a two-way analysis of variance are as follows:

1. The samples are independent, random samples from the defined populations.
2. The scores on the dependent variable are normally distributed in the population.
3. The population variances in all cells of the factorial design are equal. (Hinkle et al., 1994, p. 410)

A two-way analysis of variance test offers three advantages to the researcher. First, efficiency is a factor because the effects of two independent variables can be simultaneously investigated. Second, the researcher has the added control over variation, which enhances statistical precision. Third, the ability of the researcher to study the interaction between the independent variables is a benefit (Hinkle et al., 1994).

The interaction between the two independent variables, ability level, and mode of assessment was examined. In addition, tests for main effects for the treatments were conducted (Kirk, 1995). A retrospective power analysis and the measure of association (omega squared) was calculated for main effects.

The two-way analysis of variance allowed for the investigation of differences between the overall results for each subtest. An examination of differences on individual items from both administrative formats was also conducted. Items were examined for differences by ability group on the different formats. Item difficulty and item discrimination were calculated as well as a t-test for paired samples to determine the statistical significance between the means on the item. In addition, a second statistical method, Delta plot analysis, was also performed to detect any possible outliers between the two ability groups on the two test formats. By performing these statistical tests on individual items, it was possible to determine whether certain items performed differently based upon ability group and format.

The second research question investigated the evidence of potential item bias on the CAMT between the two ability groups used in the study. Traditionally, a study of item bias has been performed using classical item analysis. This would involve focusing on the item's difficulty and discriminating power. Both of these traditional methods of determining item bias were calculated for each item in the CAMT. Many experts in the area of item bias state that further detection of bias should also be performed (Camilli & Shepard, 1994). Hence, two additional methods for item bias detection-- the Delta plot analysis approach and differential item functioning analyses-- were used.

Jensen's Delta plot analysis is a linear transformations of the z-scores (Delta = 4z +

13). A Pearson r between the Delta Scores of different groups depicts the difference made by group membership on probability of responding correctly to the item. The higher the correlation, the less chance of item bias for that item.

The other modern measurement method used for data analysis was a logistic item response model. The item characteristic curve (ICC) for an item is considered equivalent and stable regardless of group membership (Baker, 1992). It is based on the probability of the examinee's correctly answering the question as determined by ability level. Thus, the ICC between different groups was compared. The area between the equated two curves provides an indicator of item bias. This statistical process was performed using BIMAIN software. Instead of an ICC being drawn for each item and ability group, the standard index of bias was calculated. The standard index of bias is considered to be statistically significant when the difference between the ability group item value is greater than twice the error term for that specific item (BIMAIN, 1994).

These statistical processes were performed to examine items for potential bias between subjects with different ability levels. When examining for potential item bias, it is important to be cognizant that results of statistical tests provide only an indication of item bias. It is imperative that the investigator know why the items functioned differently for the groups. Item bias can be determined only by careful study of both quantitative and qualitative data (Camilli & Shepard, 1994).

CHAPTER 4

ANALYSIS OF DATA

Introduction

The purpose of this study was to test the following research questions:

1. Does the computer-administered mathematics test produce results similar to those obtained by using the KeyMath-R, a pencil and paper test, with regard to overall scores and item statistics?

2. Do any test items in the item pool on the computer-administered mathematics test suggest differential item functioning (DIF) toward students of different mathematical ability level?

Group characteristics from the administration of both mathematics test formats are described. Findings for the study are reported by research question number as listed above.

Group Characteristics

Research participants (N=114) were administered the KeyMath-R (paper and pencil test) and the CAMT (computer test). Both tests contain 55 items. Order of the testing was determined by availability of the computer lab and scheduling by the school principal and teacher.

Subjects consistently scored higher on the KeyMath-R, which used the traditional paper and pencil format. The means and standard deviations for the four subtests from the KeyMath-R are presented in Table 7. Data from both ability groups are combined to produce the information in Table 7.

Table 7

KeyMath-R and CAMT Means and Standard Deviations

	KeyMath-R	CAMT
Subtest	Mean (SD)	Mean (SD)
Addition	10.9649 (3.6650)	8.9386 (4.0249)
Subtraction	10.7719 (4.5584)	9.2105 (4.3037)
Multiplication	10.0702 (4.4378)	8.4298 (4.7279)
Division	9.7368 (3.5798)	6.2895 (3.7763)

Cronbach's alpha based on all subjects for the KeyMath-R was .9144. Using all subjects, the CAMT had a Cronbach's alpha of .9068.

When both ability groups were examined together, variance was obtained on all items on both test formats, with the exception of problem 16 on the addition subtest of the KeyMath-R (paper and pencil version). This specific test item had no variance among either of the ability level groups (normally achieving and mathematically disabled) since no student answered the item correctly. Overall item difficulty for the KeyMath-R was .4776, with an item variance of .1274. The CAMT had an overall item difficulty of .3683 and an item variance of .1444.

Data Analysis for Research Question 1

Question 1 addresses whether there was any difference between the results obtained from using the KeyMath-R, a pencil and paper test, and the CAMT, a computer-administered test, with regard to overall scores and item statistics. Research participants were administered both formats of the mathematics test. The mathematics tests were administered during the students' designated time for math. Because all of the schools involved in the study followed different academic schedules, mathematics tests were not administered during the same time of day. However, all study participants took both tests during the morning hours. Order of the tests also varied due to the fact that not all third-grade classes could use the computer lab during their math class on the same day. The order of which classes took the computer test first was determined by the school's principal and the teacher and the availability of the computer lab. Study participants were allowed to use scratch paper regardless of the test format in order to answer the mathematic items. Unlimited time to complete each test format was given to study participants.

Analysis on the two mathematics tests was done using the scores obtained from the four subtests (addition, subtraction, multiplication, and division). A two-way analysis of variance was performed on each subtest. The independent factors were test format (paper and pencil vs. computer) and ability level (normally achieving and mathematically disabled). Interaction between the two factors was also examined. Assumptions for the two-way analysis of variance were met with the exception of equal variances in each cell of the design. Nevertheless, Hinkle et al. (1994) reported that the two-way analysis of variance is very robust to the violation of assumptions. These assumptions for the statistical test were checked by examining results gained from the descriptive statistics procedure in SAS. In addition, a t-test was performed on each item to see if the testing format affected scores. This analysis was performed according to ability group. These results were also examined by Delta plot analysis. Each subtest is discussed separately in sections below.

Addition Subtest

Initial findings for the addition subtest show that all study participants consistently scored higher results on the KeyMath-R (paper and pencil version) than on the CAMT (computer version). The mean and standard deviation for each ability group on the KeyMath-R appear in Table 8. Table 9 provides the same information concerning the CAMT.

Graphing the means of scores from the KeyMath-R and CAMT illustrates that scores on the KeyMath-R were consistently higher for both ability groups.

Table 8

KeyMath-R Means and Standard Deviation by Ability Group (Addition)

Ability group	Mean	Standard deviation
Normally achieving	12.4923	3.2745
Mathematically disabled	8.9387	3.1583

Table 9

CAMT Means and Standard Deviation by Ability Group (Addition)

Ability group	Mean	Standard deviation
Normally achieving	10.0000	3.9011
Mathematically disabled	7.4081	3.5935

A two-way analysis of variance was used to analyze both the main effects of ability group and method of assessment and also the interaction between the two factors on overall scores. There was not a statistically significant interaction between method of assessment and ability group F(1,224) = 1.05, p > .05. However, there was evidence of main effects for both ability group and method of assessment. Ability group was statistically significant F(1,224) = 42.83, p < .001, with the normally achieving group receiving higher scores. The retrospective power of the main effect for ability group was .99, with an omega square (i.e., explained variance) of .155. Method of assessment was also shown to be statistically significant F(1,224) = 18.36, p < .001, with subjects receiving higher scores on the KeyMath-R. For method of assessment, the retrospective power of the test was .98 and an omega squared of .07.

Since differences were found between methods of assessment, classical item analysis was used to examine whether differences existed between ability groups on individual items in the addition subtest. Item difficulty, item discrimination, group differences by item, and Delta plot analysis are documented in this section.

An examination of addition results by ability group revealed that the mathematically disabled group had 5 items with zero variance on the paper and pencil format (KeyMath-R). Lack of variance results when either all subjects or no subjects correctly answer the item. Item 2 had a mean of 1.0 and a standard deviation of 0.0. Items 13 through 16 all had a mean and standard deviation of 0.0. For the computer-administered version of the mathematics test, only one item had zero variance. Item 14 on the CAMT had a mean and standard deviation of 0.0.

Similar results can be seen when examining the data produced by research subjects in the normally achieving group. There were 2 items with zero variance on the paper and pencil test (KeyMath-R). Item 5 was correctly answered by all subjects and had a mean of 1.0 and a standard deviation of 0.0. Item 16 was incorrectly answered by all subjects and had a mean and standard deviation of 0.0. No items on the CAMT had a variance of zero for the normally achieving group. Additional item analysis information is provided in Appendix D.

Potential item differences between the two testing formats could also be seen in a series of t-tests and by using Delta plot analysis. For mathematically disabled participants, Items 1, 2, 3, 8, 9, 10, 11, and 16 were found to be statistically different on the two formats by ability level at the .05 level. For the listed items, all scores, with the exception of Item 16, had a higher score on the KeyMath-R than on the CAMT. Detailed information on the t-test for each item can be seen in Appendix E.

The Delta plot analysis for mathematically disabled subjects on the addition subtests reveals similar findings. Items found to function differently by t-tests were also shown to function differently using the Delta plot analysis, with one exception. Item 16 was found to be significantly different for mathematically disabled students, using the t-test analysis but not on the Delta plot analysis. It should be noted that the Delta plot analysis typically provides the investigator with a more liberal view concerning whether significant differences between the format exist. Appendix F displays the Delta plot analysis for mathematically disabled subjects on the addition subtests.

For subjects categorized as normally achieving, the t-tests revealed that Item 1 and Items 5 through 13 were statistically different at the .05 level. Item 13 is the only item from the previous list in which the score for the CAMT was higher than that obtained in the KeyMath-R. Information for each item is reported in Appendix E.

For normally achieving subjects on the addition subtests, all items found to function differently by t-test analysis were also found to function differently using Delta plot analysis. The Delta plot analysis for normally achieving students on the addition subtests is presented in Appendix G.

Subtraction Subtest

Means and standard deviations from the two ability groups establish that on the subtraction subtest all subjects scored higher on the KeyMath-R than on the CAMT. Table 10 presents the means and standard deviations of the two ability groups on the KeyMath-R, while Table 11 reports the same information for the CAMT.

Table 10

KeyMath-R Means and Standard Deviation by Ability Group (Subtraction)

Ability group	Mean	Standard deviation
Normally achieving	12.4000	4.3974
Mathematically disabled	8.6326	3.8551

Graphing the means for the KeyMath-R and the CAMT reveals that scores obtained by all study participants were higher on the KeyMath-R, regardless of ability group.

Table 11

CAMT Means and Standard Deviation by Ability Group (Subtraction)

Ability Group	Mean	Standard Deviation
Normally achieving	10.8153	4.2015
Mathematically disabled	6.9591	3.4336

A two-way analysis of variance was performed on scores from the subtraction subtest to test for both main effects for ability group and method of assessment and to test for possible interaction between the two factors. No statistically significant interaction between ability group and mode of assessment, F(1, 224) = 0.01, p > .01, was found. There were main effects for both of the independent factors in the design. Ability group was a statistically significant factor, F(1,224) = 49.87, p <.0001, with the normally achieving subjects scoring higher than the mathematically disabled subjects. The retrospective power analysis of this result was .99, with an omega squared of .17. Method of assessment also proved to be a statistically significant factor in the design F(1,224) = 9.11, p < .05. Results from the KeyMath-R were higher than those obtained on the CAMT. Retrospective power for the test was .85, and the omega squared was .03.

When item data for the subtraction subtest was examined, it was noted that no items, regardless of ability group, had a variance of zero. Appendix H presents additional item analysis for the subtraction subtests. However, results revealed a different scenario when the data are separated by ability group. For subjects who were mathematically disabled, results from t-tests on the individual items indicate that statistical differences existed between the two formats on Items 1 through 7 and Item 9 at the .05 level. With the exception of Items 2 and 3, all items had a higher score on the KeyMath-R than on the CAMT. Appendix I reports specific t-test information for each item.

The Delta plot analysis for the subtraction subtest comparing the two testing formats with mathematically disabled students concurs with the information found in the t-tests. All items found to be significant by Delta plot analysis were also found to function differently when analyzed using t-tests. The Delta plot analysis for mathematically disabled subjects for subtraction can be found in Appendix J.

T-tests for students who were normally achieving in mathematics report that Items 1, 2, 5 though 8, 10, and 14 were statistically different at the .05 level. T-test information is reported in Appendix I. All items indicating differences between formats had higher scores on the KeyMath-R, with the exception of Items 2 and 14. These two items had higher scores on the CAMT.

Delta plot analysis for normally achieving subjects on the subtraction subtests presents a similar picture, but with several exceptions. According to Delta plot analysis, Item 1 does not function differently on the two testing formats, but the t-test analysis reports that this item does function differently. Also, the t-test analysis shows Item 9 as functioning the same on the two testing formats. The Delta plot analysis clearly depicts this item as functioning differently between the two formats. The Delta plot analysis for normally achieving subjects on the subtraction subtests can be found in Appendix K.

Multiplication Subtest

Multiplication subtest data indicate that study participants, regardless of ability group, scored higher on the KeyMath-R than on the CAMT. Means and standard deviations for both test formats are reported in the Tables 12 and 13 below.

Table 12

KeyMath-R Means and Standard Deviation by Ability Group (Multiplication)

Ability group	Mean	Standard deviation
Normally achieving	12.0000	4.1193
Mathematically disabled	7.6326	3.5513

Table 13

CAMT Means and Standard Deviation by Ability Group (Multiplication)

Ability group	Mean	Standard deviation
Normally achieving	9.9692	4.8572
Mathematically disabled	6.2857	3.6055

Data from the two-way analysis of variance show that there was no statistically significant interaction between the ability groups and the mode of assessment, F(1, 224) = .38, $p > .05$. However, there were main effects for both independent factors. Ability group was statistically significant, F(1, 224) = 53.02, $p < .0001$, with the normally achieving subjects receiving higher scores. Retrospective power analysis was found to be .99, with an omega squared of .19. Mode of assessment was also statistically significant, F(1, 224) = 9.33, $p < .05$, with the higher scores reported on the KeyMath-R. The retrospective power analysis was calculated at .86, with an omega squared of .04.

Item analysis on the multiplication subtests indicate that subjects had greater difficulty in solving these problems. This can be seen by the low item difficulty on this section. Results from participants in the mathematically disabled group reveals that there were 5 items on the KeyMath-R (paper and pencil format) with zero variance. Items 9 through 13 all had a mean and standard deviation of 0.0. The difficulty encountered by the mathematically disabled subjects was also present on the CAMT. Three items, 11 through 13, had zero variation since the mean and standard deviation were both 0.0. Although research participants who were categorized as normally achieving in mathematics did have difficulty with the multiplication section, as evidenced by low item difficulty, there were no items having zero variance. Appendix L provides additional item-analysis information for the multiplication subtests.

T-tests and the Delta plot analysis for the multiplication section convey interesting observations for several items. T-tests suggest that Items 1, 3, 4, 6, and 14 were statistically significant when testing on different formats for subjects who were mathematically disabled. All of these items had higher scores on the KeyMath-R except for Item 14. The mean for Item 14 was higher on the CAMT than the KeyMath-R. Detailed information on t-test results can be found in Appendix M.

The Delta plot analysis for mathematically disabled subjects clearly demonstrates that Items 1, 3, 4, and 6 received higher scores on the KeyMath-R. Also shown on the Delta plot analysis is that Items 2, 5, and 7 function differently on the two test formats. These results were not found in the t-tests analysis. The Delta plot analysis for mathematically disabled subjects on multiplication subtests is found in Appendix N.

T-tests for normally achieving subjects indicate that Items 1, 2, 3, and 7 were statistically different on the factor of method of assessment. The data are displayed in Appendix M. All of the listed items were found to be higher on the KeyMath-R than on the CAMT. Again, this trend can be clearly seen on the Delta plot analysis for normally achieving subjects. All items found to function differently by the t-test analysis were also found to function differently on the Delta plot analysis. Appendix O contains the Delta plot analysis for normally achieving subjects on the multiplication subtests.

Division Subtest

Scores obtained from study participants on the division subtest indicate that, again, participants in both ability groups scored higher on the KeyMath-R than on the CAMT. Table 14 reports the means and standard deviations of the KeyMath-R, while Table 15 contains the means and standard deviations of the CAMT by ability groups.

Table 14

KeyMath-R Means and Standard Deviation by Ability Group (Division)

Ability group	Mean	Standard deviation
Normally achieving	10.7230	3.5465
Mathematically disabled	8.4285	3.2145

Table 15

CAMT Means and Standard Deviation by Ability Group (Division)

Ability Group	Mean	Standard Deviation
Normally achieving	7.5692	4.2203
Mathematically disabled	4.5918	2.1594

A two-way analysis of variance was performed to examine both possible main effects for ability group and method of assessment, as well as interaction between ability group and method of assessment. There was no statistically significant interaction between ability group and mode of assessment F(1, 224) = .55, p > .05. There was statistical significance found for ability group F(1, 224) = 32.64, p < .0001, with a retrospective power analysis of .99 and an omega squared of .12. On the division subtest, normally achieving subjects had higher scores than subjects who were mathematically disabled. Method of assessment also proved to be statistically significant, F(1, 224) = 57.38, p < .0001, with scores received on the KeyMath-R being higher. This statistical test had a retrospective power analysis of .99 and an omega squared of .19.

Item analysis on the division subtest indicated that study participants who were categorized as mathematically disabled again had difficulty with this mathematical concept. For this group, three of the division items on the KeyMath-R had zero variance. Items 8 through 10 had a mean and standard deviation of 0.0. There were also two items on the CAMT that students who were categorized as mathematically disabled were unable to answer correctly. Items 7 and 9 on the CAMT had a mean and standard deviation of 0.0. For subjects who were normally achieving in mathematics, no items had a variance of zero. Appendix P contains addition item-analysis results for the division subtests.

T-tests for the division subtest convey that Items 1 through 5 were statistically significant with regard to method of assessment. Both normally achieving subjects and mathematically disabled subjects received higher scores on the KeyMath-R than on the CAMT. Detailed information on the t-tests for the division section can be found in Appendix Q. The same items were found to function differently for both mathematically disabled subjects and normally achieving subjects, using Delta plot analysis and t-test analysis. The division subtests Delta plot analysis for mathematically disabled subjects can be found in Appendix R, and Appendix S contains the Delta plot analysis for normally achieving subjects.

Data Analysis for Research Question 2

Research question 2 addresses whether any test items on the computer-administered mathematics test (CAMT) suggested differential item functioning (DIF) toward students of different mathematical ability. To investigate whether any item on the CAMT functioned differently between the two ability groups, differential item functioning analyses were performed. These techniques were performed using the BIMAIN Item Maintenance Program (BIMAIN, 1994), which is a logistic item response model. Rather than plotting the item characteristic curves (ICCs) for each item from both ability groups, the standard index of bias is calculated for each item. This calculation uses information from the ICC that could have been plotted. The standard index of bias is considered significant when the differences between the group value is greater than twice the error term (BIMAIN, 1994). When significance occurs, the investigator is able to investigate possible reasons for the differences between the two groups (Camilli & Shepard, 1994). Jensen's Delta plot analysis, a classical method of detecting possible item bias, was also performed. To answer the second research question, all four subtests of the CAMT were analyzed. Findings are reported by subtest.

General Findings

For the entire CAMT, mathematically disabled subjects had an item mean of .3202 and a standard deviation of .3165. Item means ranged from .0000 to .9592. Excluding items with zero variance, the means ranged from .0204 to .9592. There was a total of six items with zero variance. Cronbach's alpha for mathematically disabled subjects on the CAMT was .8386.

Subjects who were normally achieving in mathematics had an item mean of .4308 and a standard deviation of .3008 on the CAMT. Item means ranged from .0154 to .9846. All items had variance on the CAMT for the normally achieving group. Cronbach's alpha for this group was .8995.

There were several trends seen in the data for the CAMT. First, with the normally achieving group, the first item in each section always had a lower difficulty coefficient than the second item. This was consistent across all four subtests. This trend was also evident with mathematically disabled subjects. The only exception with this group was on the division subtest, where the first and second item had the same item difficulty. Second, as would be expected on a well-ordered test, the percentage of subjects correctly responding to an item was higher at the beginning of the test than at the end. As the test progressed, the percentage of subjects correctly answering items decreased. This trend can be observed in both ability groups across all four mathematical subtests.

Addition Subtest

Subjects who were mathematically disabled encountered only one problem on which there was zero variance, Item 14. Excluding Item 14, the mean of the items was .5347. The minimum mean of .0204 occurred on Item 15 and the maximum mean of .9592 was recorded on item 4. The overall standard deviation for the items was .3535, with a Cronbach's alpha of .6728. Item discrimination indices for mathematically disabled subjects ranged from .0072 to .4937. Item difficulty and item discrimination for each item can be found in Appendix D.

Study participants who were classified as normally achieving in mathematics had no items with a variance of zero. The mean of the items was .5923. The lowest mean, .0308, was achieved on Item 16 and the highest mean, .9846, occurred on Item 2. The standard deviation for the items was .3449, with a Cronbach's alpha of .6551. Item discrimination indices ranged from -.0680 to .4313. There were 4 items (3, 4, 5, and 16) which had negative discrimination indices. Appendix D reports the item difficulty and item discrimination for each item.

Crocker and Algina (1986) report that negative discrimination occurs when item are "missed by many high-scoring examinees but are answered correctly by low-scoring examinees" (p. 314). This occurred in the addition subtest several times. The mathematically disabled ability group had 2 items with higher difficulty coefficients than the normally achieving ability group. By Crocker and Algina's (1986) definition, negative discrimination occurred on the addition subtest on Items 1 and 16.

For the addition subtest, the BIMAIN statistical program could not make adjustments for the differences between the two ability groups using the 3-parameter IRT method. Therefore, the 2-parameter IRT logistic model was used. The Logistic Ogive Model uses the 2 parameters of location and scale in its cumulative distribution functions (Baker, 1992). Statistically significant differences were found between normally achieving subjects and mathematically disabled subjects on Items 1, 8, 10, 11, 14, and 15. The standard index of bias and error term are reported in Appendix T.

Using the Delta plot analysis to examine items in the addition subtest indicates different results than the IRT method. The Delta plot analysis shows that Items 2, 9, and 10 could function differently for these two groups. Appendix U contains the Delta plot analysis for the CAMT addition subtest.

Subtraction Subtest

The item mean for mathematically disabled subjects on the subtraction subtest was .3688. The items with the lowest mean, .0204, occurred on Items 12 and 13. The highest mean of .8571 was achieved on Item 3. The standard deviation for all items was .3043. Cronbach's alpha was .6264. No items had a variance of zero for the subtraction subtest. Information on individual item difficulty and discrimination can be found in Appendix H.

For normally achieving subjects, the mean for the subtraction subtest was .5440. Item 14 had the lowest mean of .1231, whereas Item 2 had the highest mean of .8615. Standard deviation for all items was .2734. Overall item discrimination ranged from -.1773 to .6052. Items 1 and 2 had a negative discrimination index. The Cronbach's alpha coefficient for normally achieving subjects on the subtraction subtest was .6693.

According to Crocker and Algina (1986), Item 3 had a negative discrimination factor since more subjects who were mathematically disabled correctly answered the item than normally achieving subjects. Appendix H has item difficulty and item discrimination recorded for all subtraction items for normally achieving subjects.

Using the 2-parameter logistic model, the BIMAIN statistical program found no items that indicated differences between the normally achieving subjects and those who are mathematically disabled. The standard index of bias and error term for each item is reported in Appendix T. These results deviate from the ones suggested by the subtraction Delta plot analysis. This classical method suggests that Items 6, 7, and 8 function differently between the two ability groups. The Delta plot analysis is reported in Appendix V.

Multiplication Subtest

Subjects who were mathematically disabled encountered problems in the multiplication subtest. Items 11 through 13 had zero variance because all students incorrectly answered the items. Excluding the items with zero variance, the item mean was .1725. Items 9 and 10 had the lowest mean of .0204, and Item 2, with a mean of .4694, had the highest mean. The standard deviation of all items was .1627. Discrimination indices ranged from .0761 to .5494. Item difficulty and item discrimination can be found in Appendix L. Cronbach's alpha was .7536.

Normally achieving subjects also encountered difficulty but had variance on all test items for multiplication. Item mean was .3077. The minimum mean of .0462 occurred on Item 13, and the maximum mean of .6000 was on Item 2. The standard deviation for all items was .2012. Item discrimination indices ranged from .1331 to .6541. Item difficulty and discrimination can be found in Appendix L. Cronbach's alpha for all items was .8170.

Results from the BIMAIN program indicate that Items 11 through 13 functioned differently between the two ability groups. BIMAIN was able to detect differences using the 3-parameter logistic model, which uses the parameters of location, scale, and guessing (Baker, 1992). Lord refers to this third parameter as a lower bound for the item characteristic curve (Baker, 1992). Appendix T shows the standard index of bias and the error term for the multiplication subtest.

The Delta plot analysis for the CAMT multiplication subtest indicates that Items 1 through 6 potentially function differently for normally achieving subjects and mathematically disabled subjects. The Delta plot analysis is displayed in Appendix W.

Division subtest

For subjects who were categorized as mathematically disabled, there were 2 items with zero variance. Items 7 and 9 had a mean and standard deviation of 0.0. Excluding Items 7 and 9, the item mean for the division subtest was .0510. The minimum mean of .0204 was on Items 5, 6, and 8. The maximum mean of .0816 occurred on Items 1 through 3. The standard deviation for all items was .0282. Item discrimination indices range from .1211 to .5971 on division items for mathematically disabled subjects. The Cronbach's alpha was .7602. Information on item difficulty and item discrimination is reported in Appendix P.

Normally achieving subjects had a mean of .1985 for all division items. Item 9 had the minimum mean of .0154, and Item 4 had the maximum mean of .4462. Standard deviation for all division items for normally achieving subjects was .1483. Indices for item discrimination ranged from .0777 to .6995. Cronbach's alpha for the division subtest was .8178.

Analyzing results from the BIMAIN program using a 3-parameter logistic model reveals that Item 7 functions differently between normally achieving subjects and mathematically disabled subjects. Appendix T presents the standard index of bias and the error term for each item in the division subtest.

An examination of results shown in the Delta plot analysis for the CAMT division subtest suggests that Items 1 through 4 function differently between the two ability groups. Appendix X contains the Delta plot for the CAMT division subtest.

In summary, when attempting to detect items that could potentially function differently between the two ability groups, multiple statistical methods should be used (Hammill et al., 1996). To answer research question 2, Jensen's Delta plot analysis and IRT was performed on the four subtests. Jensen's Delta plot analysis, one of the classical methods for detecting potential item bias, produced different results than those found using a modern measurement method, the 2- and 3- parameter IRT technique. The discrepancies between the two results can be attributed to the fact that a Delta plot analysis will often indicate item bias if the item discrimination parameters are not equal (Crocker & Algina, 1986). Therefore, the items that have large item discrimination indices often appear to be biased. Since one group possesses more of the trait being measured (mathematical ability), the item discrimination indices are different between the two groups. More items in the Delta plot analysis appear to be biased than found using the IRT technique because ,with Jensen's Delta plot analysis, ability level is not taken into consideration, as it is with the IRT method. Nevertheless, both methods provide valuable information, which is discussed in the following chapter.

CHAPTER 5

SUMMARY OF FINDINGS, CONCLUSIONS, AND RECOMMENDATIONS

Introduction

This research study examined the problem of whether a computer-administered mathematics test can provide equivalent results for normal and mathematically disabled students while retaining similar psychometric characteristics of the paper and pencil version of the test. The overall purpose of this study was twofold. First, the viability of using a computer-administered test on elementary school children with and without mathematical disabilities was examined. Second, by investigating each item on the computer-administered mathematics test for potential bias between normally achieving and mathematically disabled populations, it was possible to determine whether certain mathematical concepts consistently distinguish between the two ability groups. This study can make several significant contributions to the fields of education and test development. It provides additional evidence of the validity of scores and inferences gained from using computer-administered tests as a method of assessment for young children instead of the traditional paper and pencil test. It indicates that only when a computer test is carefully constructed can a computer-administered test and a paper and pencil test with similar content provide equivalent and valid psychometric information. Results from the study also indicate that the computer-administered mathematics test can be used with young children who are either normally achieving in mathematics or are mathematically disabled.

The study was conducted by administering both the KeyMath-R (paper and pencil version) and the CAMT (computer-administered version) to 114 third graders from private and public schools in the Dallas area. The tests were administered during the subjects' mathematics class. Unlimited time was given to participants to complete each test, although most subjects finished each test during the time constraints of their normal mathematics class. In each testing format, subjects were allowed to use scratch paper to assist them in working the problems. Calculators, however, were not allowed for either of the testing formats.

Findings

When all analyses were completed, items that were found to be statistically different between the two testing formats (pencil and paper vs. computer-administered) and/or the two ability groups (normally achieving vs. mathematically disabled) were listed. Careful examination of the list of items was conducted in order to determine if a trend existed between items that were statistically different. Major findings are discussed below; however, some items did not fit into one of the main findings. These anomalies are discussed in the Observation section of this chapter. The findings of this research study are as follows:

Research Question 1

1. There was no statistically significant interaction between ability group (normally achieving and mathematically disabled) and mode of assessment (paper and pencil vs. computer-administered) between two mathematics tests with similar content.

2. There was statistical significance in the method of assessment used, as evidenced by scores obtained on both formats of the mathematics test. Research participants scored higher on all subtests of the paper and pencil format of the mathematics test than on the computer-administered format of the test.

3. Ability level was a statistically significant factor on both formats of the mathematics test. Subjects who were categorized as normally achieving in mathematics scored higher on all subtests of the KeyMath-R and the CAMT than subjects who were categorized as mathematically disabled.

Research Question 2

1. As indicated by CAMT items, no mathematical concepts exist that distinguish between normally achieving subjects and mathematically disabled subjects. It can be shown that, as experience with mathematical algorithms increases, the potential separation between normally achieving subjects and mathematically disabled subjects increases.

Discussion of Findings

Research Question 1: Finding 1

The first finding presented above suggests that the format in which a mathematics test is administered does not give either ability group (normally achieving or mathematically disabled) an advantage in achieving a higher score. It can be concluded from the data gained in this study that differences in presentation format for a mathematics test affect both ability levels in the same manner. In all four two-way analysis of variance tests, the interaction between ability group and method of assessment was not significant.

Research Question 1: Finding 2

Research participants in this study consistently scored higher on the traditional paper and pencil method of assessment than on the computer-administered test. This second finding was closely investigated to see why there was a trend toward lower scores on the computer-administered test. Several hypotheses are discussed in an attempt to explain why this difference between formats occurred.

First, the possibility that the testing formats (pencil and paper vs. computer-administered) do function differently when items are asked in a free response format must be considered. In many studies, such as the one conducted by Olsen et al. (1989), test items are in a multiple choice format. In the CAMT, the items are in a free response format. The manner in which the item is presented could be a significant independent variable affecting overall scores.

Second, the investigator, as well as several of the principals and teachers who participated in the administration of the tests used in this study, noted that students were not distressed by the traditional paper and pencil test. It was considered an ordinary event by the study participants, who diligently worked on the test and used numerous pages of scratch paper to complete items on the test. Testing strategies already possessed by the subjects were used to work the mathematics problems. This was not the case when subjects were taking the mathematics test on the computer. There appeared to be no strategy for working the mathematics problems on the computer-administered mathematics test.

Study participants appeared to view the computer-administered mathematics test differently from the paper and pencil version of the test. At a majority of the schools, it was thought by subject behavior that subjects perceived the computer lab as a place where they generally “played” educational programs; hence, the computer played the role of glorified educational toy. These educational programs, such as drill and practice programs, often contain games as a reward for learning/performing a skill. Several subjects expressed disappointment that there were no games to play after working numerous items on the CAMT.

Subjects were excited to get to work on the computer since all had used the computer lab many times during their academic career. One subject noted that working the problems on the computer was easier since mistakes could more easily be erased. Therefore, it appears that the format of the mathematics test generally was received enthusiastically by young children. Although enthusiasm for using the computer was high, the academic seriousness was not at the same level as that given to the paper and pencil test.

The investigator noticed that most students had no strategy for taking a test on the computer; therefore, they typed in “garbage” if they did not immediately know the answer to the mathematics problem. Many subjects quickly became frustrated and began to “just type in numbers” on the computer-administered test. When they were taking the test, the “gaming” mentality was also apparent in many subjects. It appeared as if the subjects expected a game to occur at the end of the test, so they tried in the shortest amount of time, to reach the point in the program where a game would occur. It is interesting to note that the apparent seriousness and dedication with which the participants used on the paper and pencil format was not the case for all items. On the addition subtest, Item 16 is the same on both the KeyMath-R and the CAMT. On the KeyMath-R, all subjects, regardless of ability level, incorrectly answered the item. On the CAMT, some students in both ability groups correctly answered the item. The mathematics problem was exactly the same. The only difference was the presentation format.

A concern that arose in the administration of the computer-administered mathematics test related to a lack of test taking strategy for this format was that a majority of the students would not use the scratch paper provided to assist them in working the mathematics problem. Therefore, if they could not mentally work the problem, it was not seriously attempted by most of the subjects. This alone could account for the fact that the time taken to administer the computer version was much shorter than that for the paper and pencil test.

The failure to work problems on scratch paper would contribute to the problem of simple mistakes in “carrying” used in addition and “borrowing” used in subtraction. Anomalous responses were seen in many of the addition and subtraction subtests. For normally achieving subjects, Items 6, 7, and 9 through 11 on the addition subtest and Items 5 through 8 as well as Item 10 on the subtraction subtest indicate that this could have occurred. Another possible explanation would be the manner in which students entered in their responses. Responses on Items 9, 10, and 11 from the addition subtest and 5 through 7 and Item 9 on the subtraction subtest indicate possible mistakes due to subjects who were mathematically disabled not using scratch paper.

The subjects’ not taking advantage of the opportunity to use scratch paper could also explain why such a difference in the means on the multiplication and division portion exists between the two formats. Examining the scratch paper of research participants taking the paper and pencil portion of the multiplication section reveals that students who could not multiply could still work the problem if they understood the overall concept of multiplication. For example, if the problem was 23 x 3 (Item 7 on the KeyMath-R), the student wrote down 23 three times and then added. Understanding the concept of multiplication allowed students to receive credit for many multiplication problems for which they had no algorithm to solve the item. However, this understanding of multiplication did not appear to occur on the computer-administered version of the test.

Another problem related to the presentation format of the CAMT concerns how the research participants entered their answers. Students were expected to click in the answer box and type in their answer from left to right. However, since these subjects are young and still work mathematics problems algorithmically, they tried to enter answers the same way in which they are taught to work problems. Students are given the algorithm to work from right to left (start at the ones column, then the tens column, etc.). Many students knew how to work the problems but did not receive credit for a correct answer because it was entered in reverse order. Although students were told that, by working the problems on paper and then typing in the answer they would get the answer they wanted, this was never clear to a majority of them. This could have been what made the two formats statistically different on Items 7, 9 through 11 on the addition subtest and Items 5 through 8 and 10 on the subtraction subtest for normally achieving subjects. For subjects who were mathematically disabled, this could have been the reason for Items 9 though 11 on the addition subtest and Items 5 through 7 and 9 on the subtraction subtest being statistically different according to format. This concern with the testing format would need to be corrected before the two testing formats could be considered to assess students in precisely the same manner.

In summary, the difference found between the two testing formats could be attributed to how responses are entered by the subjects, the lack of an appropriate testing strategy for a computer-administered test, subjects' perception of the computer as a testing device, a majority of the subjects disregarding the opportunity to use scratch paper to work items, and the type of item presented to the examinees.

Research Question 1: Finding 3

The third major finding, that ability level is statistically significant on the scores received by students, is to be expected. The amount of ability one has in a specific area certainly predetermines how successful one will be. Therefore, it was not unexpected for students who were normally achieving in mathematics to obtain a higher mean and a smaller standard deviation than students who were mathematically disabled. It is also important to note that this finding supports the literature's findings that computer-administered tests can consistently distinguish between normally achieving subjects and those with disabilities. This study established that the CAMT can distinguish between normally achieving subjects in mathematics and those who are mathematically disabled.

Research Question 2: Finding 1

This finding concludes that there were no specific concepts where a majority of the normally achieving subjects consistently answered the problems correctly and the mathematically disabled subjects answered incorrectly. It can be observed that, as experience with a mathematical concept increases, the separation between normally achieving subjects and mathematically disabled subjects increases. Trends can be seen in the data produced from the BIMAIN program. The addition subtest had 6 out of the 16 items indicating differences between the two ability groups. With the exception of the first problem, the normally achieving subjects had a higher mean than the mathematically disabled subjects. Item 8 was the first problem the subjects encountered that had double-digit addition without carrying. Item 10 is the first double-digit addition problem involving carrying in both the ones and tens place and Item 11 is the first item requiring the addition of three numbers. Item 14 is the first addition problem involving fractions. All mathematically disabled subjects missed this problem. Finally, Item 15 is the only problem that contains a mixture of a whole number and a decimal. In each of these cases, when a new mathematical algorithm was encountered, the mathematically disabled subjects faltered. It is hypothesized that the reason so many items in the addition subtest functioned differently between the two ability groups is that addition is the mathematical algorithm with which the subjects (who are third graders) have had the most experience.

Results from the multiplication and division subtest also support this finding. Since few third graders are familiar with these mathematical concepts, it stands to reason that few items would distinguish between the ability groups. Item means and standard deviations were low for both ability groups on the multiplication and division subtests.

The only items that were shown to function differently on the multiplication subtest were Items 11 through 13. On all three of those items, all mathematically disabled subjects answered incorrectly. Therefore, with a variance of zero for one group, the item would logically function differently even if only a small percentage of the other group answered the item correctly. Again, this was the case with Item 7 on the division subtest. This problem was the first time the subject encountered the radicand in a division problem. All mathematically disabled subjects missed this item, so again a difference between the two groups was noted. This trend conveys the idea that experience with mathematical concepts increases the separation between ability groups.

This finding can also be supported by examining the shape of the Delta plot analysis curve on each of the subtests. These graphs, found in Appendixes U through X, show similar curves for each ability group on the CAMT. It can be observed, that overall, the two ability groups miss the same type of problem. The distance between the curves indicates that the experience in dealing with certain concepts does separate the two ability groups. It should also be noted that, because there were no statistically significant differences between the two ability groups using this assessment format, there is no reason for test developers to back away from developing and standardizing computer-administered tests even if a corresponding paper and pencil test on the concept being measured does not exist.

Observations

First, for normally achieving subjects, Item 5 on the addition subtest was different on the two formats. Study of the data revealed that all subjects correctly answered the question on the KeyMath-R, whereas several subjects missed the item on the CAMT. The lack of variance on the KeyMath-R certainly is one reason why this item was found to be significantly different between formats. No other mathematical trend could be found for possible differences between the means on the two formats on this specific item.

Second, the difficulty level of some of the problems on the CAMT was higher than the corresponding problem on the KeyMath-R. Therefore, a statistical difference was found between these items. For subjects who were mathematically disabled, this happened on Item 8 of the addition subtest and Item 4 of the multiplication subtest. This situation occurred for normally achieving subjects on Item 8 of the addition subtest as well. It was noted that Item 8 on the CAMT should be reworked so that the item has the same conceptual difficulty level as the KeyMath-R test.

Third, there were several items for which there was no reasonable explanation as to why the items functioned differently between the two formats. For normally achieving subjects in mathematics, it was not possible to explain the differences on means for Item 6 on the addition subtest and Item 7 on the multiplication subtest. For mathematically disabled subjects, it was not possible to explain the differences between the means on the two formats for Item 6 on the multiplication subtest.

The fourth observation concerns the combination of pictures and corresponding sentences, which better explains the mathematical item to the subjects. Items that contain a picture and an accompanying sentence were significantly different from problems that had only pictures explaining the item. The KeyMath-R contained items with picture and an explanatory sentence, whereas the CAMT contained items using only pictures to explain the item. Items from the KeyMath-R had a higher rate of subjects passing the item that the corresponding items on the CAMT. This trend was consistent across ability groups. For normally achieving subjects, evidence of this trend can be seen for Item 1 on the addition subtest, Item 1 on the subtraction subtest, Items 1, 2, and 3 on the multiplication subtest, and the first 4 items on the division subtest. Subjects who were mathematically disabled scored higher on the KeyMath-R on Items 1 and 2 on the addition subtest, Items 1, 3, and 4 on the subtraction subtest, Item 1 on the multiplication subtest, and the first 4 items on the division subtest.

There were only two exceptions in which the use of a corresponding explanation sentence did not enhance the difficulty of the item. For one of the instances, poor choice of wording can be attributed to the low percentage of subjects passing the item. On Item 2 of the KeyMath-R subtraction subtest, the percentage of all subjects correctly answering this item is much lower than the percentage of those answering the corresponding item on the CAMT. This is due to the awkward wording of the item. The KeyMath-R item reads "This seal can balance five balls. How many more balls could it balance then?" Many subjects stated that they did not understand what the question was asking. This item tests the subjects' reading comprehension as well as their ability to subtract. It is important to have unidimensionality in the item pool in order to avoid spurious results.

The second instance in which a corresponding sentence did not assist mathematically disabled subjects in correctly interpreting an item occurred on Item 3 of the addition subtest. For this item, if the subject did not or could not carefully read the sentence, one of the numbers needed to correctly work the problem would be missed because the picture did not provide all the necessary information. Subjects who were mathematically disabled had a lower passage rate on this KeyMath-R item than on the corresponding CAMT item, where all information was available from the picture. Since this occurred only with the mathematically disabled group, it is possible that problems with reading could have greatly inhibited this group in correctly responding to the item.

This observation reveals that combining a verbal sentence along with a picture better delineates the problem to the subject. This is only true, however, if the sentence used in the item is well worded and if the picture provides all the information needed to solve the problem.

Fifth, the alignment of the decimal point in a mathematical item affected subject response. Traditionally, decimals are aligned directly beneath each other so that a straight line is formed. This allows for the decimal to be kept in the correct place while the problem is being solved. Typically, the only time this heuristic is not followed is to determine whether a student understands how to work problems involving decimals. On the KeyMath-R addition subtest, the decimals are not aligned in Item 13. The corresponding problem on the CAMT has the decimals properly aligned. A higher percentage of normally achieving subjects correctly answered the CAMT item than the KeyMath-R item. Where the CAMT item addressed only the concept of addition, the KeyMath-R item addressed concepts of addition and the understanding of how decimals work. Traditionally, decimals are introduced to students after the fundamental concepts of addition, subtraction, multiplication, and division have been mastered. It is interesting to note that for items involving addition of decimals, Items 12 and 13, the difficulty indexes for all subjects were very low. However, the items were significantly different between formats for normally achieving subjects only. It is as if the subjects who were mathematically disabled completely ignored the decimal point in the problem. Therefore, for young children such as those used in this study, addition and decimals are separate concepts and should be assessed independently.

The sixth observation is one that would be expected. Items involving "advanced" concepts will separate ability groups and signify subjects who truly have an understanding of the overall mathematical concept. Multiplication and division are concepts that are introduced in the later part of third grade. However, this test was administered during the fall term of the academic school year. During this period of time, addition and subtraction skills are typically reinforced. Multiplication and division are usually introduced during the second half, or spring semester, of the school year. All teachers reported that multiplication and division had not officially been presented to the students. Administering the multiplication and division subtests of the KeyMath-R and CAMT in the fall semester essentially acted as a pretest for subjects. It is not surprising that the item difficulty indexes were so low and that many items had zero variance. Results would have been even lower if all items had only a mathematical problem instead of problems with corresponding pictures. The first four items on both the multiplication and division test had a picture that coincided with the problem. These pictures allowed some students to correctly answer the four items.

Problems involving fractions also posed problems for many subjects. When the subjects encountered a problem such as 1/2 - 1/3, many subjects reported that this problem was impossible to work. The understanding of subtraction was evident, but not in conjunction with an understanding of what 1/2 or 1/3 truly meant. For mathematically disabled subjects, lack of understanding about fractions was noticeable in Item 16 on the addition subtest and Item 14 on the multiplication subtest. This trend was seen in Item 14 on the subtraction subtest for subjects who were normally achieving in mathematics.

In conclusion, although these observations about the data found in this study do not directly answer one of the research questions, the topics certainly merit discussion and could lead to an avenue of future research.

Conclusions

Based upon the data gained from the study and the methodology used, the following conclusions can be drawn.

1. A well-constructed and designed computer-administered mathematics test can assess young elementary aged subjects as well as a traditional paper and pencil mathematics test. It was concluded that this did not occur in this study due to reasons discussed in finding 2. The magnitude of the correlation coefficient between the testing formats (paper and pencil vs. computer-administered) could vary depending upon the type of item format used in the computer-administered test.

2. A well-constructed and designed computer-administered mathematics test can distinguish between subjects who are mathematically disabled or normally achieving in mathematics. This supports the use of computer-administered tests with special populations.

3. Testing strategies used by young children on paper and pencil tests do not necessarily transfer to computer-administered testing situations. Subjects from both ability groups demonstrated the lack of testing strategies on the computer-administered test.

4. Normally achieving subjects in mathematics tend to recognize the decimal as a significant element in mathematical problems. Subjects who are mathematically disabled are inclined to disregard the decimal and attempt to work the problem as a whole number.

5. The manner in which subjects enter their response is a significant factor on any computer-administered mathematics test. The way to enter responses must match the traditional approach students use to work the problem on pencil and paper. These conclusions can provide test developers with insight on producing exceptional computer-administered tests for young children.

Test developers who carefully construct computer-administered mathematical tests can duplicate results found in the review of literature showing that no significant differences exist between a computer-administered test and a paper and pencil test. Results in this study found significant differences between the testing formats. Possible reasons for these differences have been discussed above. It is believed that the main reasons were the type of test item used and the entering of the examinee's response. It is paramount that the test developer allow subjects to enter their responses into the computer from right to left. Without this important factor, young children such as those used in this study do not necessarily make the "conceptual leap" that they must enter their responses into the answer box from left to right. This will reduce the probability of significant differences in the method of assessment. Thus, having examinees enter responses in the same order as they do on a pencil and paper test will strengthen the argument that the two testing formats function in the same manner.

By making the computer a trivial tool in the testing process and not allowing this tool to determine how responses must be entered, results gained from computer-administered mathematics tests will be statistically the same as those found from the traditional paper and pencil mathematics test.

Test developers should also take note that the type of item presented to young children could significantly alter the correlation coefficients between an existing paper and pencil test and a newly developed computer-administered test. Studies with a high correlation coefficient between a paper and pencil test and a computer-administered test present the items in a multiple choice format (Olsen et al., 1989; Olsen, 1990). The McDonald et al. (1992) study, containing free response items, had a lower correlation coefficient when comparing pencil and paper test with a computer-administered test.

In addition, by carefully constructing a unidimensional item pool, the computer-administered test can consistently distinguish between subjects who are normally achieving in mathematics and those who are mathematically disabled. This was shown consistently in results from the CAMT in that subjects who were normally achieving had higher means than those obtained by the mathematically disabled subjects.

This study also indicates that different ability groups approach certain topics in varied ways. Decimals is a concept in which the difference in confronting the problem was apparent between normally achieving subjects and mathematically disabled subjects.

It always behooves the researcher to investigate whether the statistical differences found on certain topics are equivalent to "real world" significance. In this case, the differences between formats should be considered. The differences found in this study would most likely diminish if subject responses were entered from right to left. Format differences would also decrease if subjects were taught several testing strategies that could be used on a computer-administered mathematics exam.

Recommendations and Implications

The implications of this study point to the fact that computer-administered tests can be used with young children and special populations. Again, the test must be well developed and needs to match the response technique used by young children to accomplish the desired task. This study also provides additional evidence that a well designed computer-administered test can distinguish between subjects with and without learning disabilities. Using computer-administered tests in the educational system also allows the examinee and test administrator to benefit from the advantages associated with this testing paradigm. The use of computerized testing will certainly extend beyond the educational system. For example, many prison systems test inmates, and the use of computer-administered tests could be used in this situation.

More research on computerized testing should be performed using subjects with various disabilities. There are many studies on the use of the computer and writing skills with special populations. Nevertheless, the areas of math and science have little research on computer-administered testing with special populations.

In addition, a replication of this study which expanded the subject population, both in numbers, type of school, and addition of grade levels, would be beneficial. Increasing the sample for this study would greatly increase the ability to generalize the results found in the expanded study. Although this research study contained subjects from both the private and public school sectors, the majority of subjects in the study were from private schools. Because several schools in this study focused on learning disabilities, subjects could have been better trained in taking tests than students with a learning disability in the public school.

Additional research can also be conducted in the area of mathematics. It would be extremely beneficial for mathematics teachers to know exactly where the conceptual understanding between normally achieving students and mathematical disabled students exists. This information would allow teachers to focus efforts on these concepts and thus allow greater achievement on the part of the student with mathematical disabilities.

Finally, studies could be performed on the different strategies students use to work a computer-administered test. Subjects must be taught testing strategies for computer-administered mathematics tests. The strategies used by young subjects on traditional paper and pencil tests do not appear to be assimilated as a technique for the new testing format. Related studies have been performed on college students who have taken the computerized Graduate Records Exam (GRE). As the Preliminary Scholastic Aptitude Test (PSAT) and Scholastic Aptitude Test (SAT) are computerized, one can assume that additional studies of this nature would be performed. However, it is critical that subjects not always be at the high school level or above. The inclusion of young children in this type of study is important.

Finally, although computer-administered testing offers considerable potential, there is more to test development than simply changing the medium of presentation. It is clear that expressions, knowledge, ability, type of item used, and presentation all affect student performance and must be considered in test development.

APPENDIX A

KEYMATH-R ITEMS

KEYMATH-R ITEMS

Addition

1. 3 + 1 *

2. 5 + 2 *

3. 4 + 2 *

4. 2 + 1 *

5. 3 + 5

6. 9 + 6

7. 26 + 50

8. 18 + 5

9. 81 + 45

10. 34 + 69

11. 261 + 40 + 715

12. $137.01 + 87.45

13. 26.3 + 15.472

14. 2/5 + 1 /5

15. 1.6 + 2

16. 3/4 + 1/8

Subtraction

1. 5 - 2

2. 5 - 4

3. 5 - 3

4. 4 - 1

5. 9 - 2

6. 16 - 7

7. 98 - 30

8. 62 - 5

9. 73 - 29

10. 285 - 187

11. 500 - 304

12. $40.00 - 29.25

13. 7/9 - 2/9

14. 1/3 - 1/4

Multiplication

1. 4 x 4

2. 4 x 5

3. 4 x 6

4. 9 x 0

5. 3 x 7

6. 6 x 8

7. 23 x 3

8. 20 x 7

9. 47 x 6

10. 502 x 8

11. $40.27 x 3

12. 83 x 20

13. 49 x 23

14. 1/2 x 3/4

15. 7 x 1/6

Division

1 . 10 ÷ 4

2. 12 ÷ 2

3. 18 ÷ 3

4. 15 ÷ 3

5. 21 ÷ 3

6. 54 ÷ 6

7. $3 \sqrt{96}$

8. $5 \sqrt{85}$

9. $6 \sqrt{540}$

10. $20 \sqrt{820}$

APPENDIX B

CAMT ITEMS

CAMT ITEMS

Addition

1. 2 + 1 *
2. 6 + 3 *
3. 4 + 3 *
4. 3 + 3 *
5. 6 + 5
6. 8 + 6
7. 28 + 6
8. 25 + 34
9. 64 + 55
10. 84 + 37
11. 123 + 20 + 861
12. $278.22 + 72.63
13. 18.3 + 26.691
14. 2/7 + 3 /7
15. 2.8 + 3
16. 3/4 + 1/8

Subtraction

1. 3 - 1
2. 5 - 2
3. 7 - 3

4. 4 - 0

5. 12 - 5

6. 17 - 9

7. 67 - 20

8. 53 - 5

9. 81 - 17

10. 175 - 79

11. 900 - 463

12. $60.00 - 29.25

13. 3/4 - 1/4

14. 1/2 - 1/3

Multiplication

1. 3 x 2

2. 4 x 5

3. 3 x 8

4. 9 x 5

5. 2 x 0

6. 6 x 7

7. 32 x 3

8. 30 x 6

9. 63 x 8

10. 203 x 9

11. $20.95 x 4

12. 72 x 50

13. 83 x 29

14. 1/2 x 3/5

15. 9 x 1/5

Division

1. 6 ÷ 2

2. 12 ÷ 3

3. 16 ÷ 2

4. 15 ÷ 5

5. 32 ÷ 4

6. 72 ÷ 9

7. $2\sqrt{86}$

8. $5\sqrt{95}$

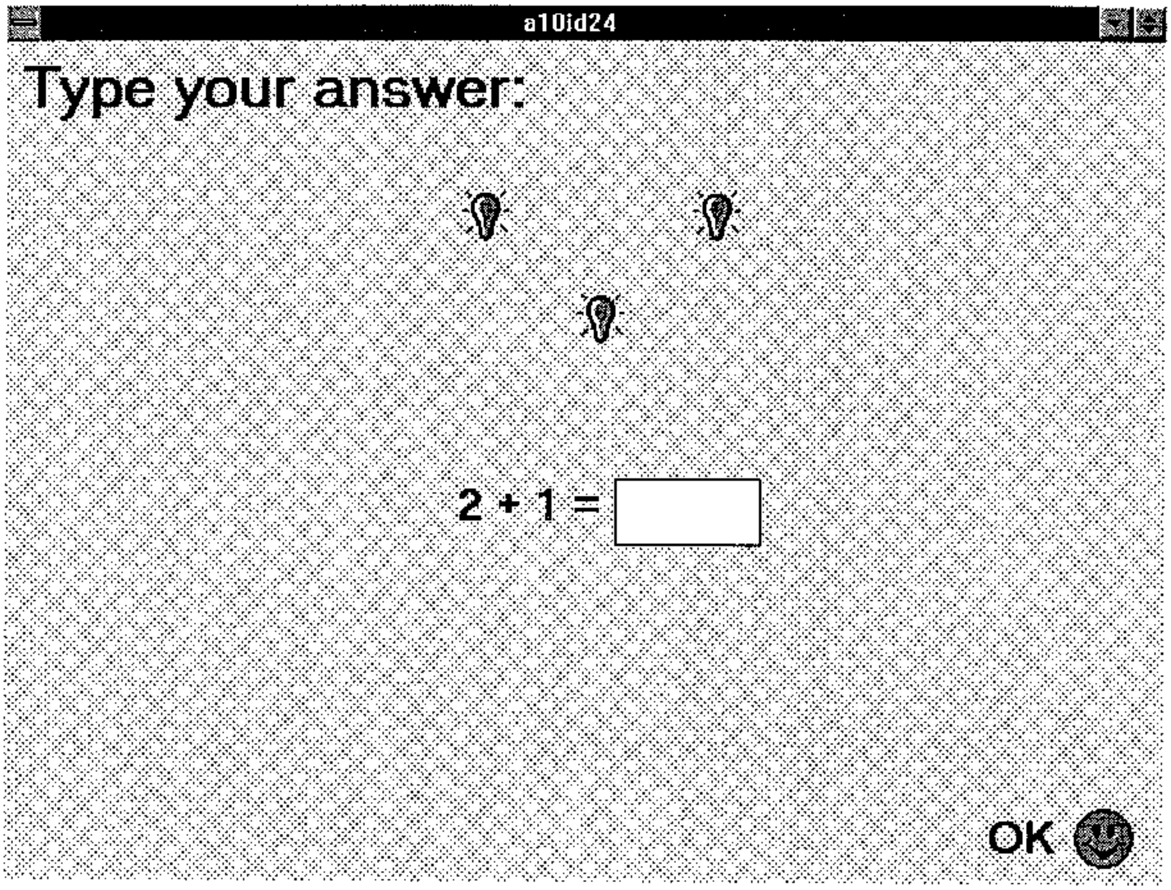
9. $4\sqrt{120}$

10. $30\sqrt{960}$

APPENDIX C

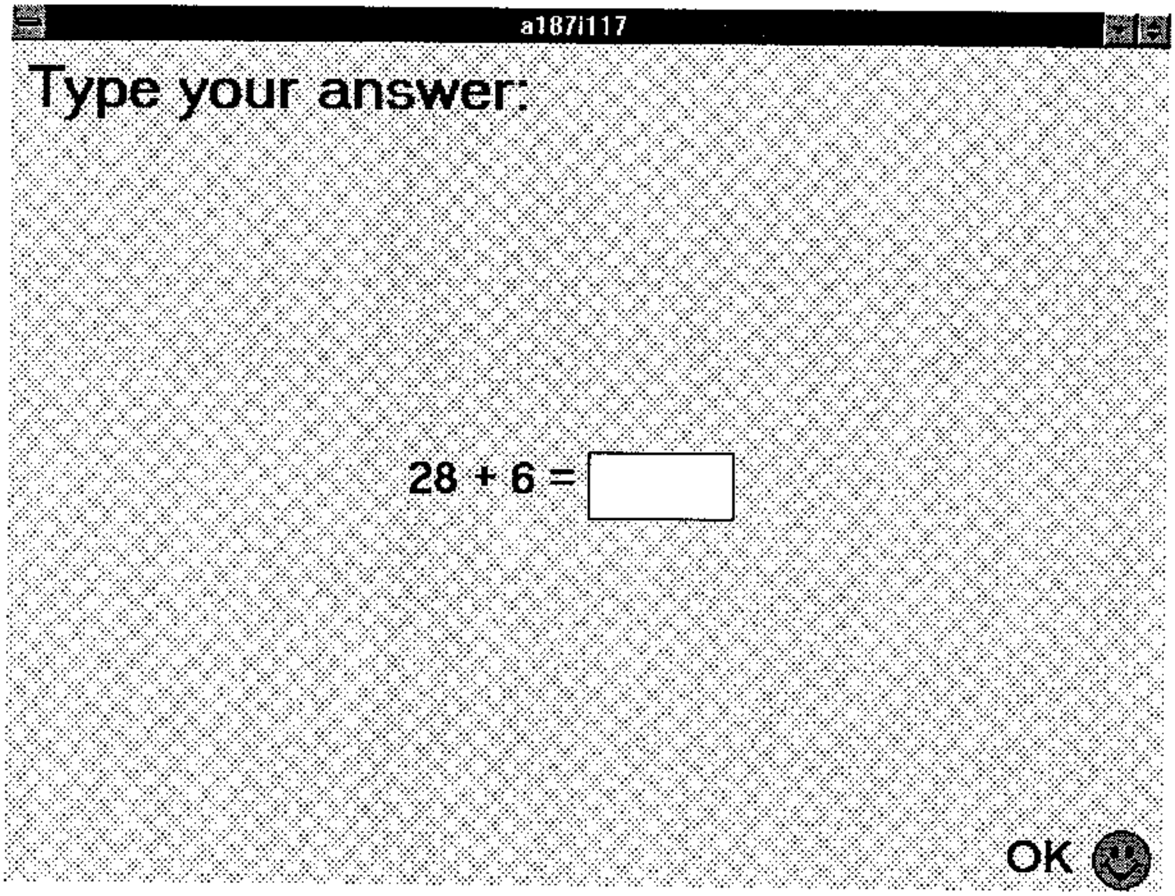
SAMPLE SCREENS FOR CAMT

SAMPLE SCREENS FOR CAMT



Type your answer:

2 + 1 =



Type your answer:

28 + 6 = □

APPENDIX D

ITEM ANALYSIS FOR ADDITION

Table 16

Item Analysis for Addition (Mathematically Disabled)

Item	Item difficulty	Item discrimination
1	.7551	.1280
2	.8571	.2155
3	.9184	.3322
4	.9592	.2264
5	.8776	.3770
6	.8367	.3039
7	.8571	.2365
8	.5102	.1921
9	.4898	.4937
10	.3673	.3366
11	.3265	.4190
12	.1224	.2380
13	.0408	.2305
14	.0000	-------
15	.0204	.0587
16	.0816	.0072

Note. Dashes indicate analysis could not be performed.

Table 17

Item Analysis for Addition (Normally Achieving)

Item	Item difficulty	Item discrimination
1	.6615	.0980
2	.9846	.0103
3	.9692	-.0680
4	.9692	-.0680
5	.9385	-.0148
6	.8462	.3404
7	.8769	.3911
8	.7077	.3406
9	.7692	.3021
10	.5692	.4065
11	.3692	.2004
12	.2769	.2790
13	.3077	.4313
14	.0769	.1959
15	.1231	.4075
16	.0308	-.0034

APPENDIX E

T-TESTS ANALYSIS FOR ADDITION SUBTEST

Table 18

T-tests for Mathematically Disabled Subjects

Item	df	t-value	p
1	48	-2.64	.011*
2	48	-2.83	.007*
3	48	3.79	.000*
4	48	-.57	.569
5	48	-1.94	.058
6	48	-1.95	.057
7	48	-.70	.485
8	48	-3.65	.001*
9	48	-2.86	.006*
10	48	-2.14	.038*
11	48	-2.07	.044*
12	48	-.70	.485
13	48	1.43	.159
14	NA	NA	NA **
15	48	1.00	.322
16	48	2.07	.044*

Note. An asterisk denotes a significant finding at the .05 level.

Double asterisks denotes the standard error of the difference is 0 and the analysis cannot be performed.

Table 19

T-tests for Normally Achieving Subjects

Item	df	t-value	p
1	64	-5.17	.000*
2	64	.00	1.000
3	64	1.43	.159
4	64	.81	.418
5	64	-2.05	.045*
6	64	-2.78	.007*
7	64	-2.42	.018*
8	64	-3.96	.000*
9	64	-2.61	.011*
10	64	-4.63	.000*
11	64	-6.22	.000*
12	64	-3.01	.004*
13	64	5.14	.000*
14	64	.38	.709
15	64	.00	1.000
16	64	1.43	.159

Note. An asterisk denotes a significant finding at the .05 level.

APPENDIX F

DELTA PLOT ANALYSIS FOR ADDITION (MATHEMATICALLY DISABLED)

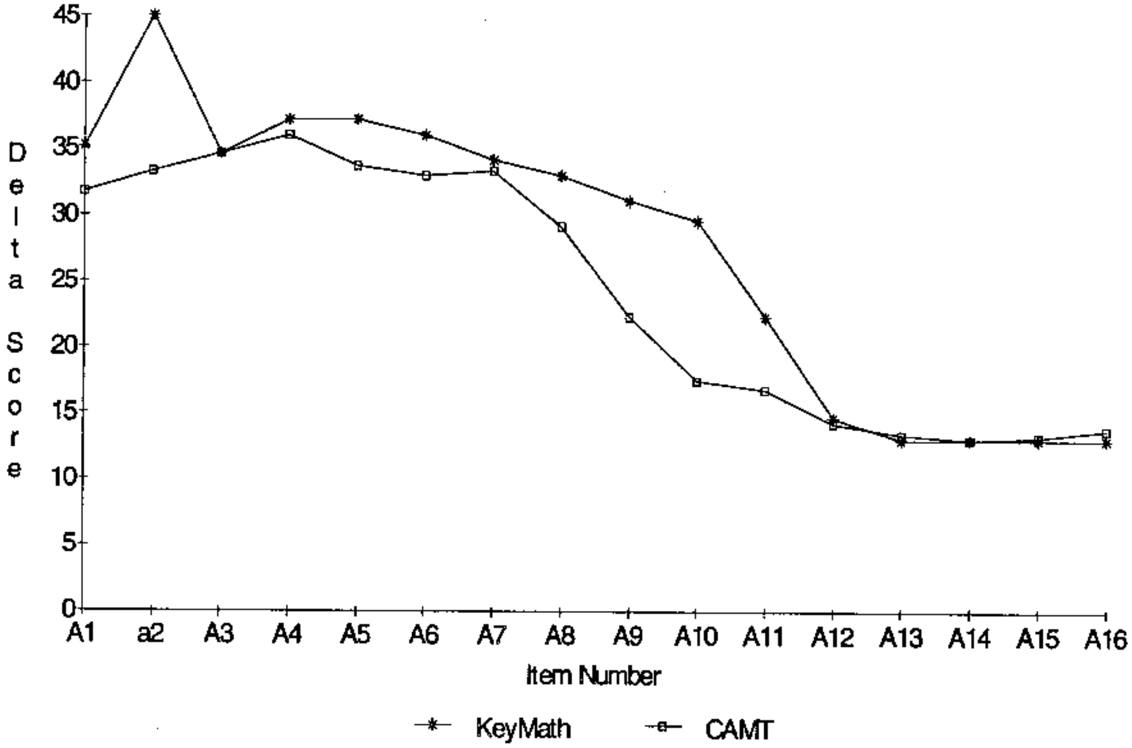


Figure 1. Delta plot analysis for mathematically disabled subjects on addition subtests.

APPENDIX G

DELTA PLOT ANALYSIS FOR ADDITION (NORMALLY ACHIEVING)

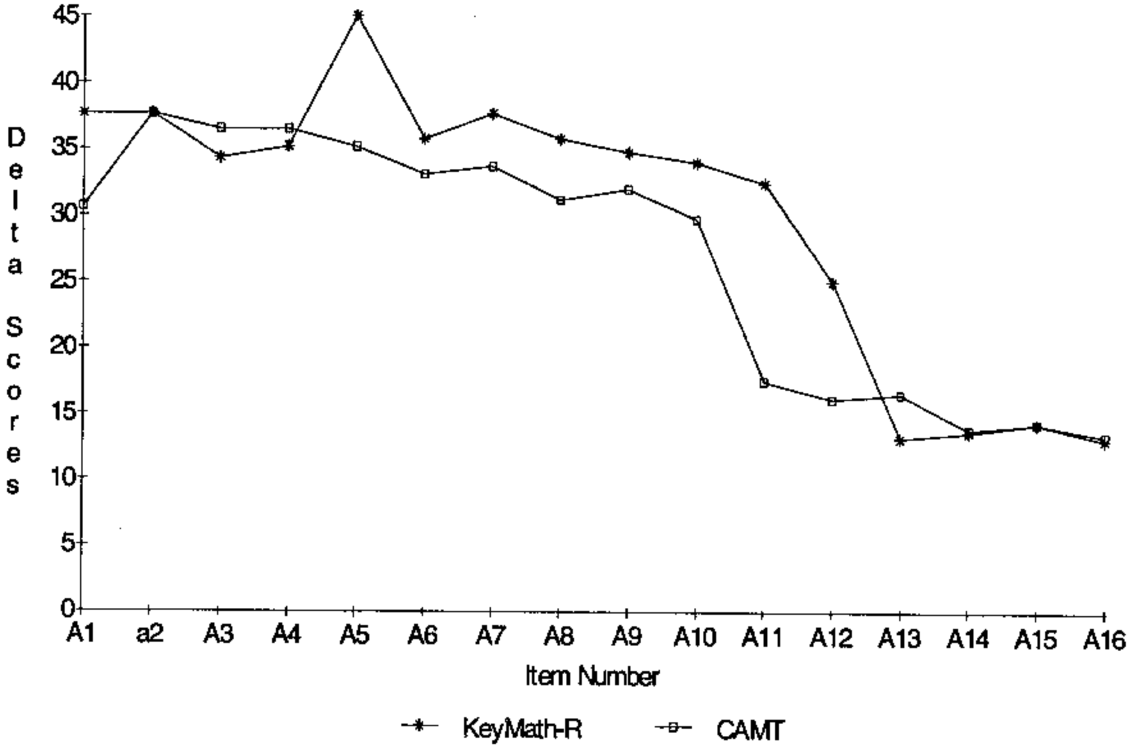


Figure 2. Delta plot analysis for normally achieving subjects on addition subtests.

APPENDIX H

ITEM ANALYSIS FOR SUBTRACTION SUBTESTS

Table 20

Item Analysis for Subtraction (Mathematically Disabled)

Item	Item difficulty	Item discrimination
1	.6531	-.0098
2	.8367	.2041
3	.8571	.2260
4	.6122	.3364
5	.5714	.1878
6	.4490	.3913
7	.4694	.4925
8	.2653	.3700
9	.1633	.2429
10	.1429	.4201
11	.0612	.4408
12	.0204	.4930
13	.0204	.0587
14	.0408	.0667

Table 21

Item analysis for Subtraction (Normally Achieving)

Item	Item difficulty	Item discrimination
1	.7385	-.1773
2	.8615	-.0856
3	.8308	.0503
4	.8308	.2101
5	.8308	.3080
6	.7231	.3993
7	.7077	.4808
8	.4769	.3950
9	.4308	.4722
10	.3385	.6052
11	.3077	.5339
12	.2769	.5472
13	.1385	.3159
14	.1231	.2117

APPENDIX I

T-TESTS ANALYSIS FOR SUBTRACTION SUBTESTS

Table 22

Subtraction T-tests for Mathematically Disabled Subjects

Item	df	t-value	p
1	48	-4.82	.000*
2	48	5.32	.000*
3	48	2.88	.006*
4	48	-2.40	.020*
5	48	-4.33	.000*
6	48	-3.27	.002*
7	48	-3.26	.002*
8	48	-1.30	.200
9	48	-2.69	.010*
10	48	-.90	.371
11	48	-1.66	.103
12	48	-1.35	.182
13	48	.00	1.000
14	48	.57	.569

Note. An asterisk denotes a significant finding at the .05 level.

Table 23

Subtraction T-tests for Normally Achieving Subjects

Item	df	t-value	p
1	64	-4.05	.000*
2	64	4.07	.000*
3	64	-1.07	.289
4	64	-.77	.443
5	64	-3.21	.002*
6	64	-3.59	.001*
7	64	-3.21	.002*
8	64	-2.55	.013*
9	64	-1.94	.057
10	64	-3.37	.001*
11	64	-1.84	.070
12	64	-.26	.799
13	64	1.69	.096
14	64	2.42	.018*

Note. An asterisk denotes a significant finding at the .05 level.

APPENDIX J

DELTA PLOT ANALYSIS FOR SUBTRACTION (MATHEMATICALLY DISABLED)

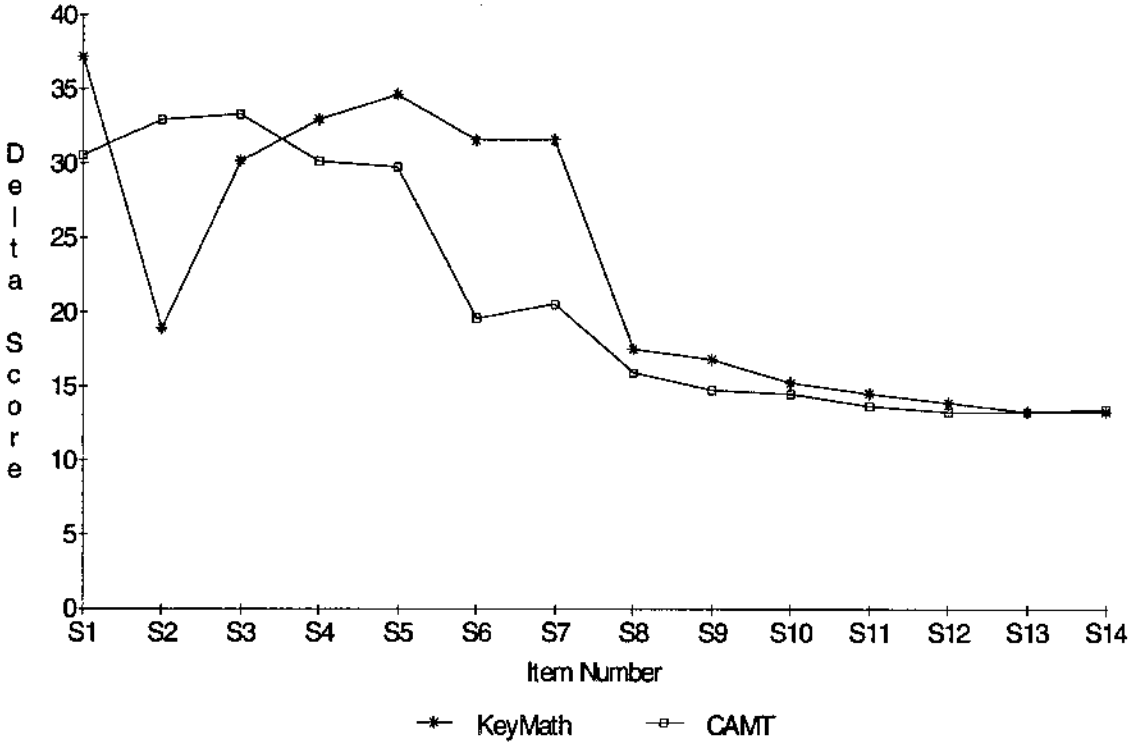


Figure 3. Delta plot analysis for mathematically disabled subjects on subtraction subtests.

APPENDIX K

DELTA PLOT ANALYSIS FOR SUBTRACTION (NORMALLY ACHIEVING)

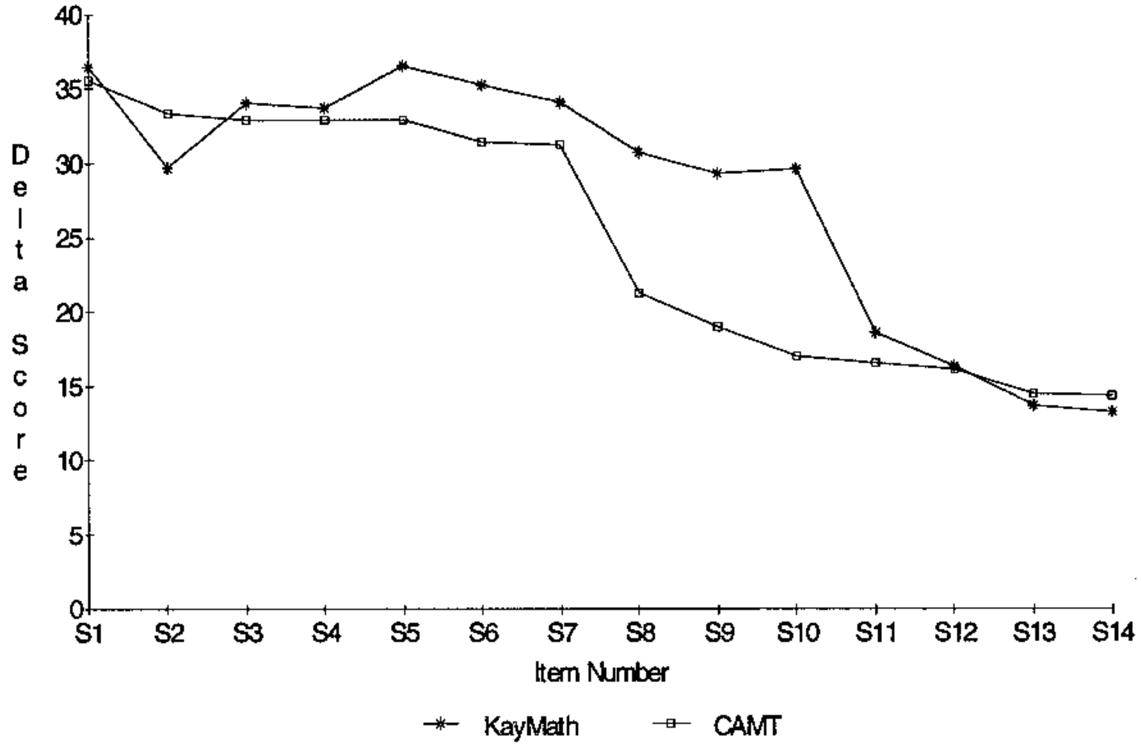


Figure 4. Delta plot analysis for normally achieving subjects on subtraction subtests.

APPENDIX L

ITEM ANALYSIS FOR MULTIPLICATION SUBTESTS

Table 24

Item Analysis for Multiplication (Mathematically Disabled)

Item	Item difficulty	Item discrimination
1	.3878	.1500
2	.4694	.3535
3	.3673	.3839
4	.1837	.3726
5	.3061	.3525
6	.0408	.4890
7	.0816	.2979
8	.0612	.5494
9	.0204	.4930
10	.0204	.4930
11	.0000	-------
12	.0000	-------
13	.0000	-------
14	.1020	.1071
15	.0612	.0761

Note. Dashes indicated analysis cannot be performed.

Table 25

Item Analysis for Multiplication (Normally Achieving)

Item	Item difficulty	Item discrimination
1	.4923	.3282
2	.6000	.2696
3	.5077	.4611
4	.5231	.5890
5	.5846	.6466
6	.3692	.3968
7	.3691	.6495
8	.3385	.6541
9	.2308	.6516
10	.1538	.3874
11	.1077	.3750
12	.0769	.3879
13	.0462	.1722
14	.1538	.1331
15	.0615	.2820

APPENDIX M

T-TESTS ANALYSIS FOR MULTIPLICATION SUBTESTS

Table 26

Multiplication T-tests for Mathematically Disabled Subjects

Item	df	t-value	p
1	48	-2.22	.032*
2	48	-1.59	.118
3	48	-3.27	.002*
4	48	-2.69	.010*
5	48	-1.77	.083
6	48	-2.07	.044*
7	48	-1.94	.058
8	48	-.44	.659
9	48	1.00	.322
10	48	1.00	.322
11	NA	NA	NA**
12	NA	NA	NA**
13	NA	NA	NA**
14	48	2.07	.044*
15	48	1.00	.322

Note. An asterisk denotes a significant finding at the .05 level.

Double asterisks denotes the standard error of the difference is 0 and the analysis cannot be performed.

Table 27

Multiplication T-tests for Normally Achieving Subjects

Item	df	t-value	p
1	64	-5.18	.000*
2	64	-2.67	.009*
3	64	-4.51	.000*
4	64	-1.15	.254
5	64	-1.72	.090
6	64	-.72	.471
7	64	-2.25	.028*
8	64	-1.15	.254
9	64	1.40	.167
10	64	-1.07	.289
11	64	-1.40	.167
12	64	1.65	.103
13	64	.44	.658
14	64	.00	1.000
15	64	-.33	.742

Note. An asterisk denotes a significant finding at the .05 level.

APPENDIX N

DELTA PLOT ANALYSIS FOR MULTIPLICATION (MATHEMATICALLY DISABLED)

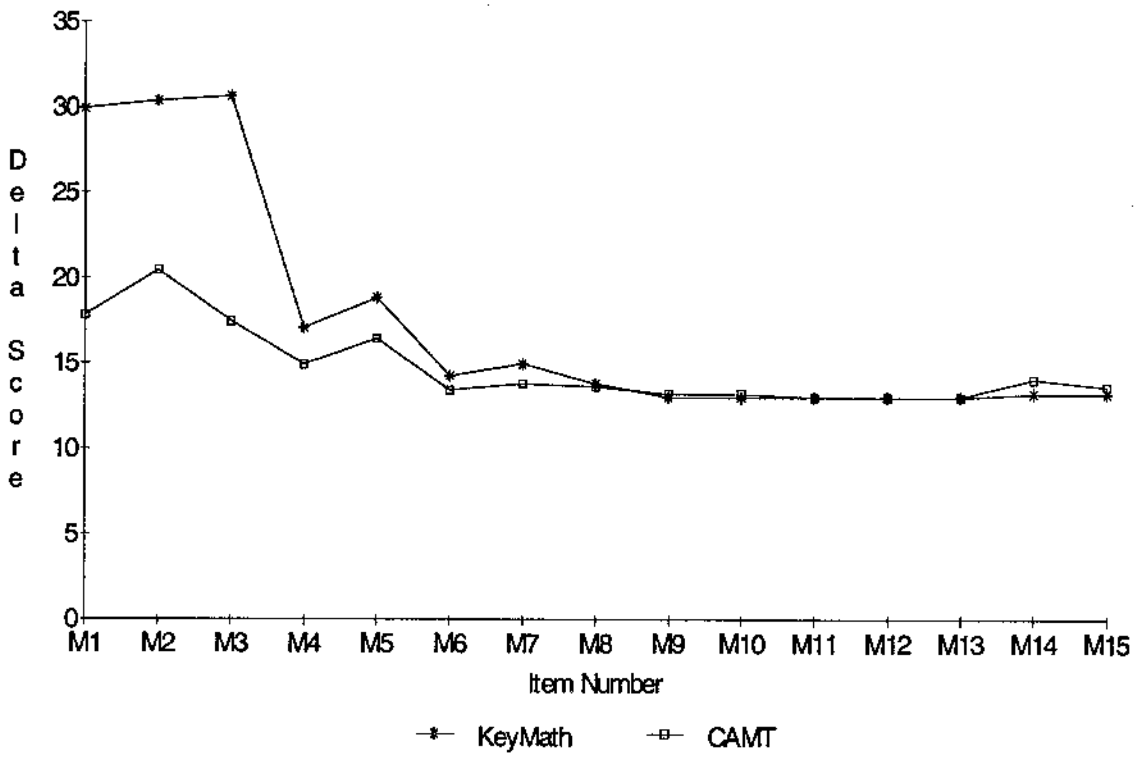


Figure 5. Delta plot analysis for mathematically disabled subjects on multiplication subtests.

APPENDIX O

DELTA PLOT ANALYSIS FOR MULTIPLICATION (NORMALLY ACHIEVING)

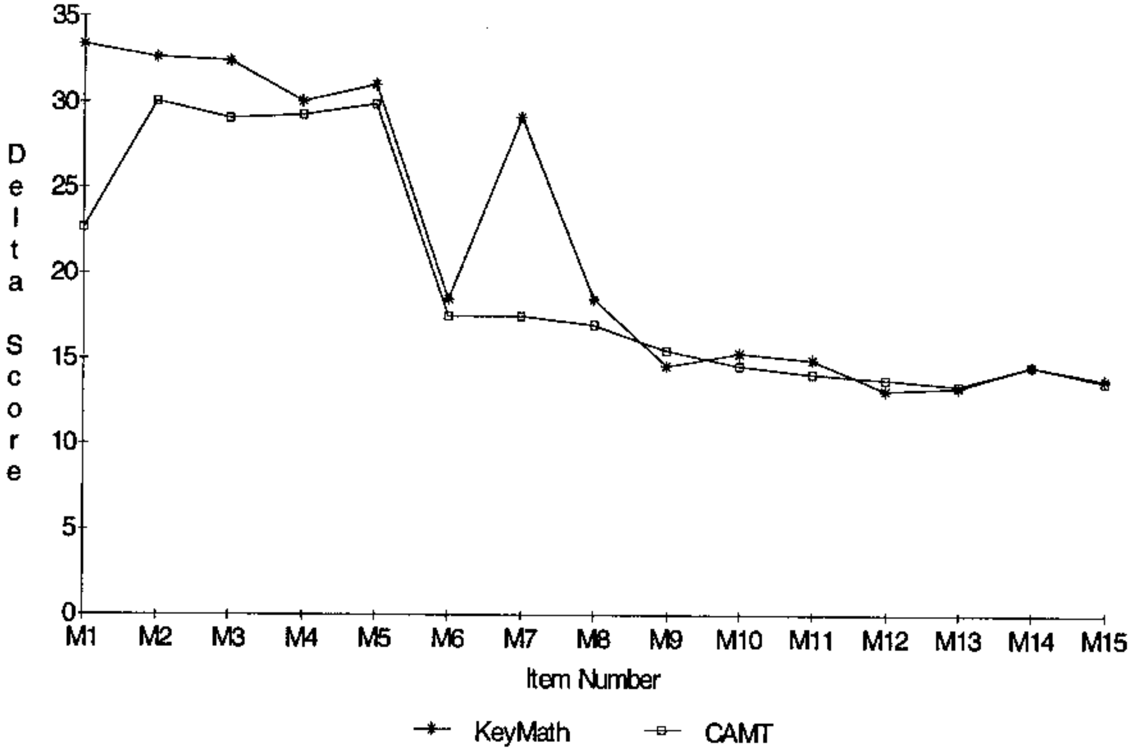


Figure 6. Delta plot analysis for normally achieving subjects on multiplication subtests.

APPENDIX P

ITEM ANALYSIS FOR DIVISION SUBTESTS

Table 28

Item Analysis for Division (Mathematically Disabled)

Item	Item difficulty	Item discrimination
1	.0816	.3652
2	.0816	.3517
3	.0816	.5971
4	.0612	.5494
5	.0204	.4930
6	.0204	.4930
7	.0000	-------
8	.0204	.4930
9	.0000	-------
10	.0408	.1211

Note. Dashes indicate analysis could not be performed.

Table 29

Item Analysis for Division (Normally Achieving)

Item	Item difficulty	Item discrimination
1	.3231	.5593
2	.3385	.5284
3	.3231	.6166
4	.4462	.6995
5	.1538	.4234
6	.1692	.4753
7	.0769	.3948
8	.0923	.3983
9	.0154	.0777
10	.0462	.3802

APPENDIX Q

T-TESTS ANALYSIS FOR DIVISION SUBTESTS

Table 30

Division T-tests for Mathematically Disabled Subjects

Item	df	t-value	p
1	48	-6.14	.000*
2	48	-6.56	.000*
3	48	-6.83	.000*
4	48	-5.79	.000*
5	48	-2.34	.024*
6	48	-1.77	.083
7	48	-1.00	.322
8	48	1.00	.322
9	NA	NA	NA**
10	48	1.43	.159

Note. An asterisk denotes a significant finding at the .05 level.

Double asterisks denotes the standard error of the difference is 0 and the analysis cannot be performed.

Table 31

Division T-tests for Normally Achieving Subjects

Item	df	t-value	p
1	64	-6.63	.000*
2	64	-6.22	.000*
3	64	-6.63	.000*
4	64	-3.03	.003*
5	64	-3.00	.004*
6	64	.28	.784
7	64	-.38	.709
8	64	.38	.709
9	64	-.57	.568
10	64	1.00	.321

Note. An asterisk denotes a significant finding at the .05 level.

APPENDIX R

DELTA PLOT ANALYSIS FOR DIVISION (MATHEMATICALLY DISABLED)

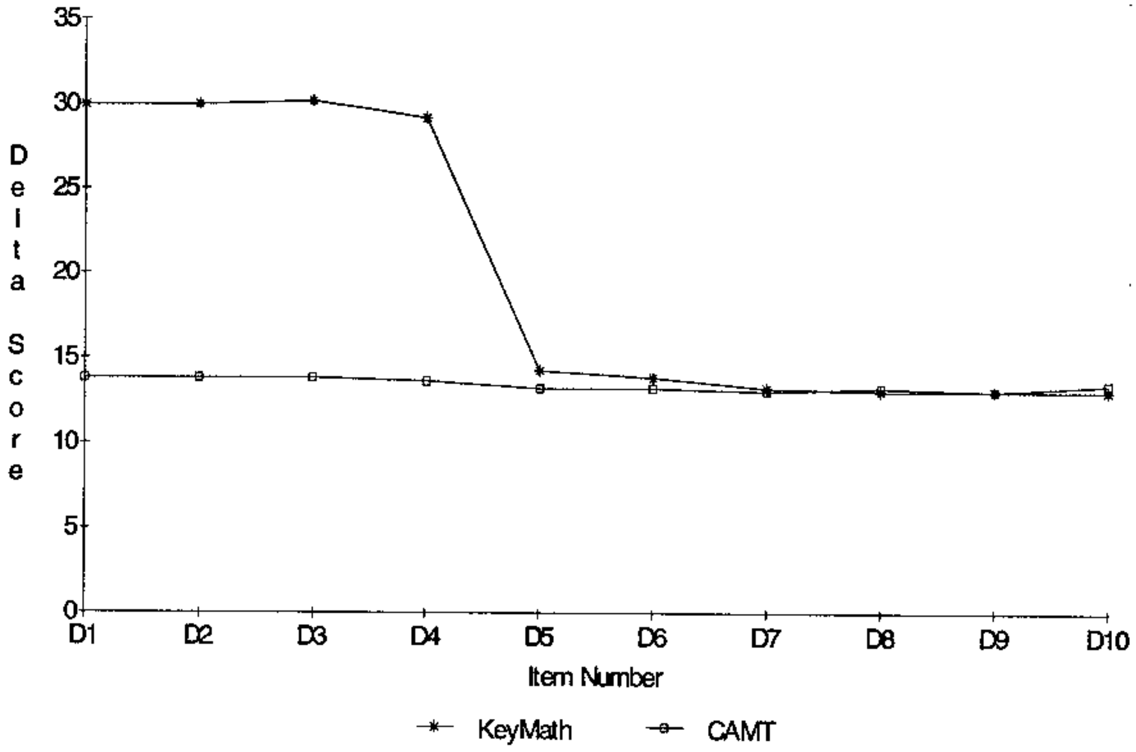


Figure 7. Delta plot analysis for mathematically disabled subjects on division subtests.

APPENDIX S

DELTA PLOT ANALYSIS FOR DIVISION (NORMALLY ACHIEVING)

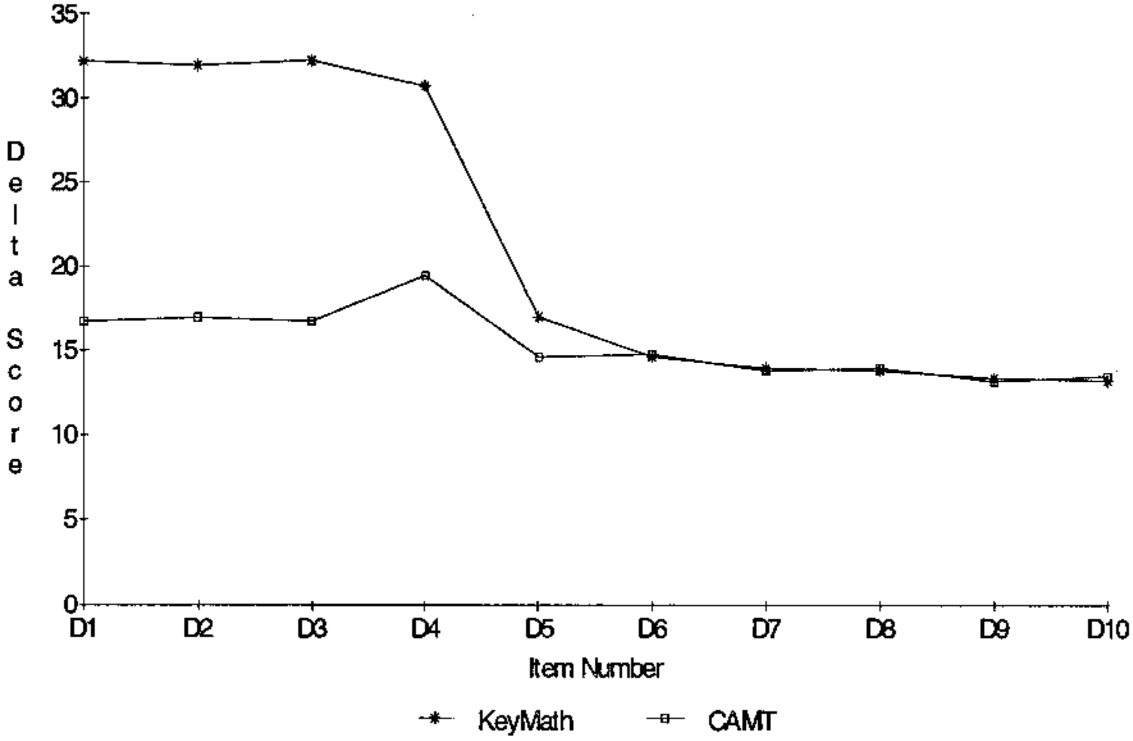


Figure 8. Delta plot analysis for normally achieving subjects on division subtests.

APPENDIX T

BIMAIN ANALYSIS FOR CAMT SUBTESTS

Table 32

Addition Subtest

Item	Standard index of bias	Error
1	4.378	1.993
2	1.442	2.484
3	0.940	2.570
4	2.548	2.470
5	0.821	2.613
6	2.444	1.631
7	2.158	2.075
8	1.239	0.450
9	0.620	0.489
10	1.294	0.426
11	2.225	0.618
12	1.515	0.785
13	0.568	0.969
14	3.254	0.684
15	24.457	0.954
16	5.149	11.708

Table 33

Subtraction Subtest

Item	Standard index of bias	Error
1	0.624	2.316
2	0.669	2.508
3	1.703	1.817
4	0.274	0.668
5	0.428	0.610
6	0.175	0.391
7	0.551	0.285
8	0.397	0.396
9	0.147	0.326
10	0.406	0.343
11	0.462	0.495
12	1.011	0.521
13	0.831	1.214
14	0.418	1.617

Table 34

Multiplication Subtest

Item	Standard index of bias	Error
1	0.382	0.308
2	0.350	0.403
3	0.276	0.276
4	0.281	0.181
5	0.102	0.144
6	0.866	0.440
7	0.282	0.219
8	0.065	0.359
9	0.143	0.224
10	0.029	0.255
11	1.043	0.170
12	1.124	0.205
13	1.752	0.428
14	0.249	1.350
15	0.510	1.238

Table 35

Division Subtest

Item	Standard index of bias	Error
1	0.082	0.226
2	0.157	0.262
3	0.098	0.169
4	0.333	0.195
5	0.136	0.488
6	0.137	0.491
7	1.317	0.230
8	0.083	0.823
9	--------	--------
10	0.664	0.712

Note. Dashes indicated analysis could not be calculated.

APPENDIX U

CAMT ADDITION DELTA PLOT ANALYSIS

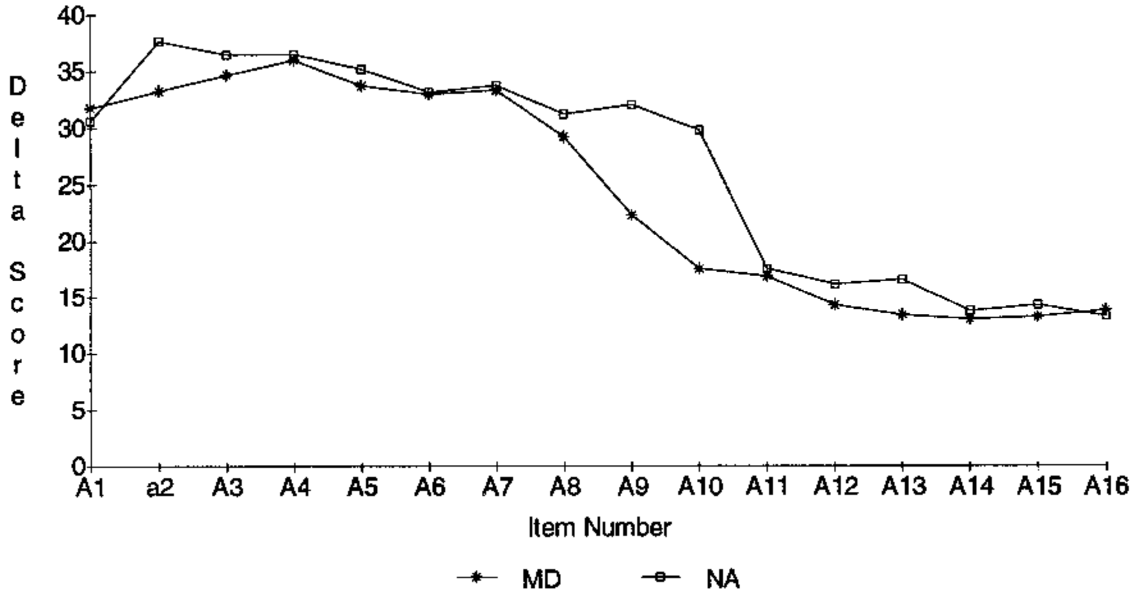


Figure 9. CAMT addition Delta plot analysis.

APPENDIX V

CAMT SUBTRACTION DELTA PLOT ANALYSIS

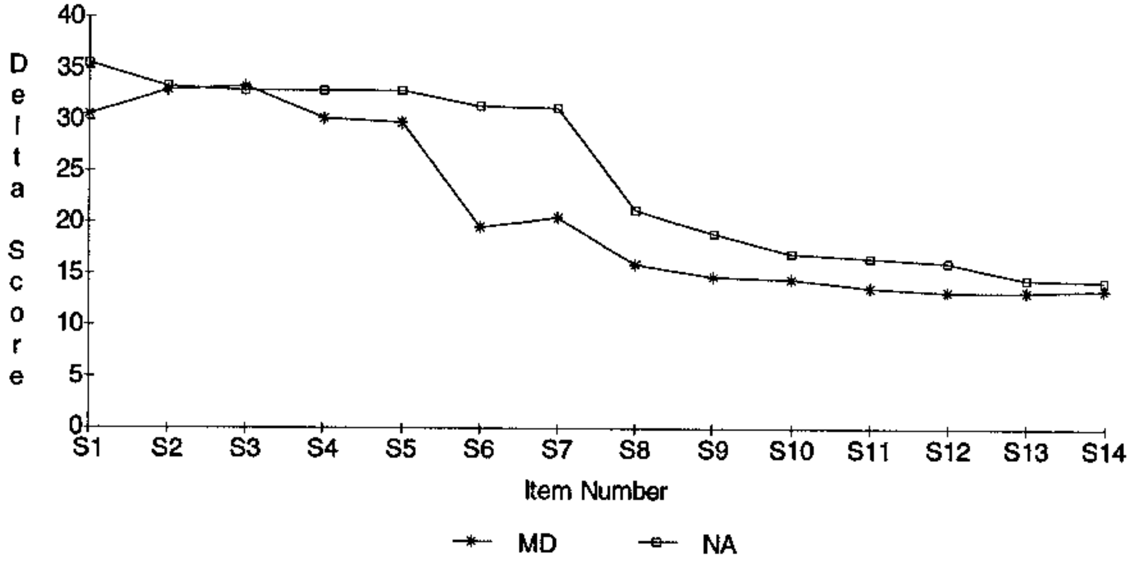


Figure 10. CAMT subtraction Delta plot analysis.

APPENDIX W

CAMT MULTIPLICATION DELTA PLOT ANALYSIS

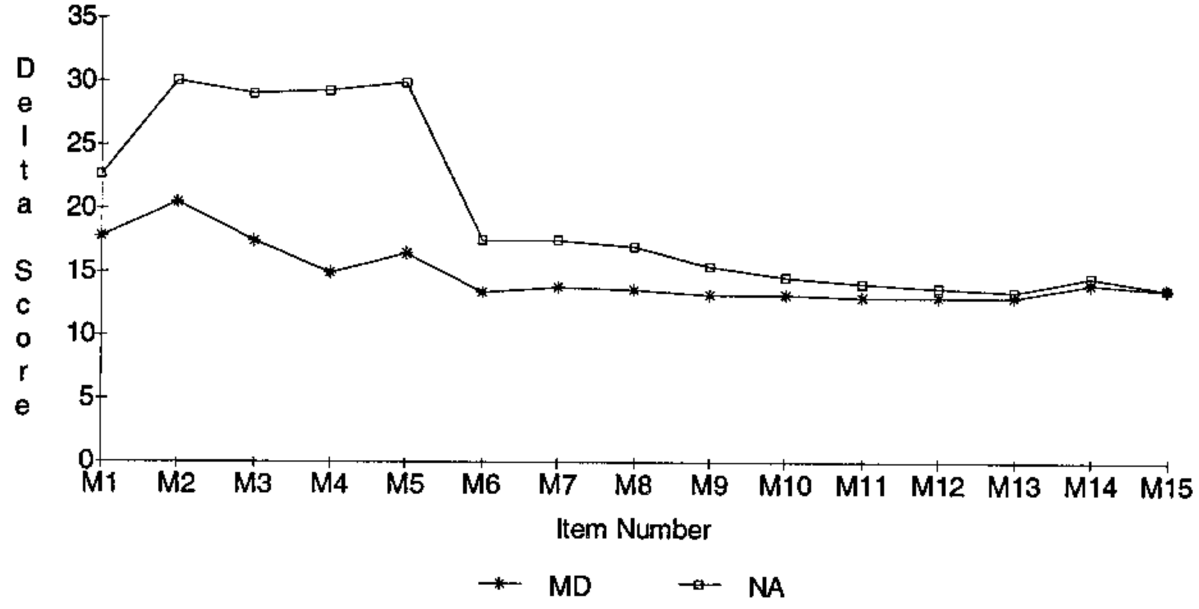


Figure 11. CAMT multiplication Delta plot analysis.

APPENDIX X

CAMT DIVISION DELTA PLOT ANALYSIS

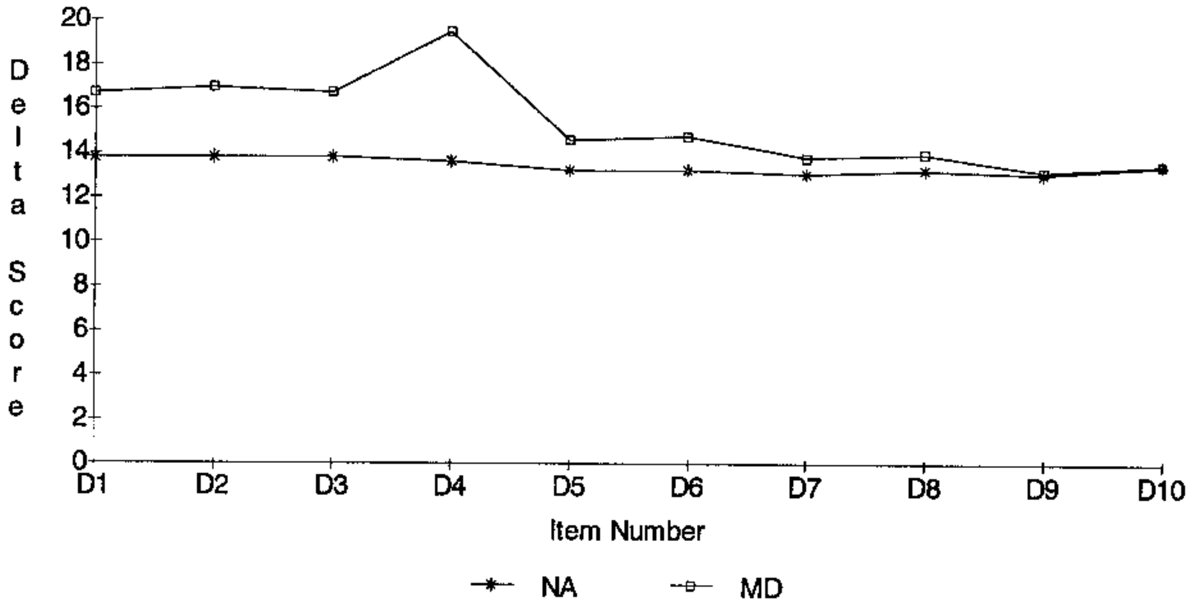


Figure 12. CAMT division Delta plot analysis.

REFERENCES

Angoff, W.H. (1988). Validity: An evolving concept. In H.Wainer & H.I. Braun (Eds.), Test validity (pp. 19-30). Hillsdale, NJ: Erlbaum.

Allen, C.C., Ellinwood, E.H., & Logue, P.E. (1993). Construct validity of a new computer-assisted cognitive neuromotor assessment battery in normal and inpatient psychiatric samples. Journal of Clinical Psychology, 49(6), 874-882.

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

American Psychological Association, American Educational Research Association and National Council on Measurement in Education. (1985). Standard for educational and psychological testing. Washington, DC: American Psychological Association.

Baker, F. B. (1992). Item response theory: Parameter estimation techniques. New York: Marcel Kekker.

Berger, S.G., Chibnall, J.T., & Gfeller, J.D. (1994). The category test: A comparison of computerized and standard versions. Assessment, 1(3), 255-258.

Bergstrom, B.A. (1992, April). Ability measure equivalence of a computer adaptive and pencil and paper tests: A research synthesis. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.

Bergstrom, B.A., Lunz, M.E., & Gershon, R.C. (1992). Altering the level of difficulty in computer adaptive testing. Applied Measurement in Education, 5, 137-149.

BIMAIN Item Maintenance Program: Logisitic Item Response Model [Computer software]. (1994). Chicago: Scientific Software International.

Bronson, M. B. (1985, April). Developing computerized assessment in young chidren. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago.

Brown, F.G. (1983). Principles of educational psychology (3rd ed.). New York: Holt, Rinehart & Winston.

Bugbee, A.C., & Bernt, F.M. (1990). Testing by computer: Findings in six years of use 1982-1988. Journal of Research on Computing in Education, 23(1), 87-100.

Camilli, G., & Shepard, L.A. (1994). Methods for identifying biased test items. Thousand Oaks,CA: Sage.

Campbell, D.T., & Stanley, J.C. (1963). Experimental and quasi-experimental designs for research. Boston: Houghton Mifflin.

Carlson, R.D. (1994). Computer adaptive testing: A shift in the evaluation paradigm. Journal of Educational Technology Systems, 22, 213-224.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Erlbaum.

Connolly, A.J. (1988). KeyMath revised: A diagnostic inventory of essential mathematics manual forms a and b. Circle Pines, MN: American Guidance Service.

Crocker, L.M., & Algina, J. (1986). Introduction to classical and modern test theory. Fort Worth, TX: Harcourt Brace Jovanovich College Publishers.

Cronbach, L.J. (1988). Five perspectives on validity argument. In H. Wainer &

H.I. Braun (Eds.), Test validity (pp. 3-17). Hillsdale: Erlbaum.

DeAyala, R.J. (1992). The influence of dimensionality on cat ability estimation. Educational and Psychological Measurement, 31, 251-263.

Dimcock, P.H., & Cormier, P. (1991). The effects of format differences and computer experience on performance and anxiety on a computer-administered test. Measurement and Evaluation in Counseling and Development, 24, 119-126.

Dorans, N.J., & Schmitt, A.P. (1991). Constructed response and differential item functioning: A pragmatic approach (College Board Report No. ETS--RR-91-47). Princeton, NJ: Educational Testing Service.

Engdahl, B. (1991, August). Computerized adaptive assessment of cognitive abilities among disabled adults. Paper presented at the meeting of the American Psychology Association Annual Convention, San Francisco.

Feinstein, Z. S. (1995). Effects of differing item parameters on closed-interval DIF statistics. Applied Psychological Measurement, 19(2), 131-142.

Ferguson, G.A. (1981). Statistical analysis in psychology and education (5th ed.). New York: McGraw-Hill.

Fletcher, P., & Collins, M.A.J. (1986-1987). Computer-administered versus written tests- advantages and disadvantages. Journal of Computers in Mathematics and Science Teaching, 6(2), 38-43.

Fuchs, L.S., Fuchs, D., & Hamlett, C.L. (1992). Computer applications to facilitate curriculum-based measurement. Teaching Exceptional Children, 24, 58-60.

Green, B.F. (1988). Construct validity of computer-based tests. In H. Wainer &

H.I. Braun (Eds.), Test validity, (pp. 77-86). Hillsdale, NJ: Erlbaum.

Gronlund, N.E., & Linn, R.L. (1990). Measurement and evaluation in teaching (6th ed.). NewYork: Macmillan.

Groth-Marnat, G., & Schumaker, J. (1989). Computer-based psychological testing: Issues and guidelines. American Orthopsychiatric Association, 59(2), 257-263.

Hammill, D.D., Pearson, N.A., & Wiederholt, J.L. (1996). Comprehensive test of nonverbal intelligence. Austin, TX: PRO-ED.

Henly, S.J., Klebe, K.J., McBride, J.R., & Cudeck, R. (1989). Adaptive and conventional versions of the DAT: The first complete test battery comparison. Applied Psychological Measurement, 13, 363-371.

Hinkle, D.E., Wiersma, W., & Jurs, S.G. (1994). Applied statistics for the behavioral sciences (3rd ed.). Boston: Houghton Mifflin.

Hsu, T., & Shermis, M.D. (1989). The development and evaluation of a microcomputerized adaptive placement testing system for college mathematics. Journal of Educational Computing Research, 5, 473-485.

HyperCard [Computer Software]. (1987). Cupertino, CA: Apple.

Katz, L., & Dalby, J.T. (1981). Computer-assisted and traditional psychological assessment of elementary-school aged children. Contemporary Educational Psychology, 6, 314-322.

Kirk, R.E. (1995). Experimental design: Procedures for the behavioral sciences (3rd ed.). Pacific Grove, CA: Brooks/Cole.

Kraemer, H.C., & Thiemann, S. (1987). How many subjects? Statistical power analysis in research. Newbury Park, CA: Sage.

Kumar, D.D., & Helgeson, S.L. (1995). Trends in computer applications in science assessment. Journal of Science Education and Technology, 4(1), 29-36.

Law, K. S. (1995). The use of Fisher's Z in Schmidt-Hunter-type meta-analyses. Journal of Educational and Behavioral Statistics, 20(3), 287-306.

Legg, S.M., & Buhr, D.C. (1992). Computerized adaptive testing with different groups. Educational Measurement, Issues and Practice, 11(2), 23-27.

Lunz, M.E., & Bergstrom, B.A. (1994). An empirical study of computerized adaptive test administration conditions. Journal of Educational Measurement, 31, 251-263.

Lunz, M.E., Bergstrom, B.A., & Wright, B.D. (1992). The effect of review on student ability and test efficiency for computer adaptive tests. Applied Psychological Measurement, 16(1), 33-40.

McDonald, J., Beal, J., & Ayers, F. (1987). Computer administered testing: Diagnosis of addition computational skills in children. Journal of Computers in Mathematics and Science Teaching, 7, 38-43.

McDonald, J., Beal, J., & Ayers, F. (1992). Details of performance on computer and paper administered versions of a test of whole number computation skills. Focus on Learning Problems in Mathematics, 14, 15-27.

Moreno, K.E., Wetzel, C.D., McBride, J.R., & Weiss, D.J. (1984). Relationship between corresponding Armed Services Aptitude Battery (ASVAB) and computerized adaptive testing (CAT) subtest. Applied Psychological Measurement, 8, 155-163.

Noijons, J. (1994). Testing computer assisted language testing: Towards a checklist for calt. CALICO, 12, 37-58.

Olsen, J.B. (1990). Applying computerized adaptive testing in schools. Measurement and Evaluation in Counseling and Development, 23, 31-38.

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

Olsen, J.B., Cox, A., Price, C., Strozeski, M., & Vela, I. (1990). Development, implementation, and validation of a computerized test for statewide assessment. Educational Measurement: Issues and Practice,9(2), 7-10.

Pressman, E., Roche, D., Davey, J., & Firestone, P. (1986). Patterns of auditory perception skills in chidren with learning disabilities: A computer-assisted approach. Journal of Learning Disabilities, 19(8), 485-488.

Reckase, M.D. (1989). Adaptive testing: The evolution of a good idea. Educational Measurement, Issues and Practice, 8(3), 11-15.

Salvia, J., & Ysseldyke, J.E. (1995). Assessment in special and remedial education (6th ed.). Boston: Houghton Mifflin.

SAS for Windows. [Computer software]. (1994).Cary, NC: SAS Institute, Inc.

Sattler, J.M. (1992). Assessment of children (3rd ed.). San Diego, CA: Jerome M. Sattler.

Shneiderman, B. (1992). Designing the user interface (2nd ed.). Reading, MA: Addison-Wesley.

Signer, B.R. (1991). CAI and at-risk minority urban high school students. Journal of Research on Computing in Education, 24(2), 189-203.

SPSS for Windows. [Computer software]. (1995). Chicago: SPSS, Inc.

Steinberg, L., Thissen, D., & Wainer, H. (1990). Validity. In H. Wainer (Ed.), Computerized adaptive testing: A primer (pp. 187-227). Hillsdale, NJ: Erlbaum.

Stocking, M.L. (1987). Two simulated feasibility studies in computerized adaptive testing. Applied Psychology: An International Review, 36, 263-277.

Texas Education Agency. (1996, September). State Board of Education rules and regulations, Title 19, Part II, Chap. 89 (Tex Reg 7240). Austin: State of Texas Printing Office.

U. S. Congress, Office of Technology Assessment. (1992, February). Testing in American schools: Asking the right questions (QTA-SET-519). Washington: U.S. Government Printing Office.

Vispoel, W.P., & Coffman, D. (1992). Computerized adaptive testing of music-related skills. Bulletin of the Council for Research in Music Education, 112, 29-49.

Visual Basic 3.0 [Computer software]. (1995). Redmond, WA: Microsoft Corportation.

Wainer, H. (1990). Computerized adaptive testing: A primer. Hillsdale, NJ: Erlbaum.

Wepner, S.B. (1991). On the cutting edge with computerized assessment. Journal of Reading, 35, 62-65.

Wilson, S., Thompson, J.A., & Wylie, G. (1982). Automated psychological testing for the severely physically handicapped. International Journal of Man-Machine Studies, 17, 291-296.

Windows 3.11 [Computer software]. (1993). Redmond, WA: Microsoft Corporation.

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., & Plake, B.S. (1990). Computer-based testing in higher eduction. Measurement and Evaluation in Counseling and Development, 23(4), 3-10.