8605066

Wisniewski, Dennis Russell

AN APPLICATION OF THE RASCH MODEL TO COMPUTERIZED ADAPTIVE TESTING: THE BINARY SEARCH METHOD

Wayne State University

PH.D. 1985

University
Microfilms
International 300 N. Zeeb Road, Ann Arbor, MI 48106

AN APPLICATION OF THE RASCH MODEL TO COMPUTERIZED ADAPTIVE TESTING: THE BINARY SEARCH METHOD

by

DENNIS RUSSELL WISNIEWSKI

A DISSERTATION

Submitted to the Graduate School
of Wayne State University,
Detroit, Michigan
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY
1985

MAJOR: EVALUATION AND RESEARCH (EDUCATION)

Approved by:

4/24/85

Adviser date

Acknowledgements

A study such as this and its findings are the responsibility of the author. However, the author's efforts alone would not be entirely sufficient for a successful completion of the study. The assistance, guidance, and suggestions by several individuals have aided and comforted the author in completion of it. The author wishes to thank the many individuals who directly or indirectly assisted in the completion of this work.

I wish to extend by deep gratification to Dr. Donald Marcotte, my major advisor, for his numerous suggestions, guidance, and concern to the significant details of this study.

Further thanks are in order to my committee members, Dr. Guy Doyal, Dr. Eugene P. Smith, Dr. T.C. Sun, and Dr. P. Hsiao for their time and careful criticism which is deeply appreciated.

Considerable gratitude and appreciation is offered to Mr. Michael Ponder of the Center for Research, Evaluation and Information Services of Pontiac Schools for his involvement in this study. It was with his assistance in many of the details of this study that its completion was realized.

I wish to thank Mr. Alfred Pavlish and the data processing department of the Center for their assistance with data entry and programming. Special thanks go to Ms. Sally Hon and Ms. Brenda Westfall for their fine efforts.

I offer my appreciation to Mr. Edward Slawski and Dr. William Veitch for their suggestions involving Rasch model theory and its relation to this study.

Many thanks go to Dr. Larry Cates for his fine efforts in proofreading the manuscript and technical assistance.

Foremost, I wish to thank my wife, Diane, for the patience, understanding and wisdom which she gave so unselfishly during these past years of study.

Finally, I thank my sons, Kurt and Mark, for having the good nature to tolerate my frequent absence from them over these years of academic endeavor.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

BACKGROUND OF THE PROBLEM

Introduction

Perhaps no invention in history has made as rapid an impact on people's lives as the computer. According to one writer, if manned flight had developed as quickly as has computer technology, "an astronaut could have orbited the earth nine years after Orville Wright wobbled aloft at Kitty Hawk."

The application of computers in business, industry, and education continues to grow at an accelerated rate. Every person comes into contact with computers in some way. Because today's students live in a computerized world, they must learn to deal with these machines that have created the "Second Industrial Revolution."

Education has been affected by developments in computer technology. However, there has been a considerable difference between currently available computer technology and the era of technology employed in education. In short, there has been a reluctance to implement the computer into the school curricula, and changes in the implemenation of technology in the educational arena have not yet arrived. This inertia will be removed if the cost of printed matter continues its exponential increase.

One area of education which has demonstrated a definitive need for computerization is testing. The computer can provide the items to the examinee, analyze the responses, and score the test according to the student's individual ability. It provides this information virtually at the touch of a key, immediately upon completion of the testing situation. Such immediate feedback is vital to any form of instruction. Accordingly, several researchers have studied this and other related aspects (Reckase 1981, Weiss 1982, Green 1982).

The argument in favor of individualized testing has been advanced by many testing and educational theorists. Notable among these is B. F. Skinner. Skinner (1968) stated, "Failure to provide for differences among students is perhaps the greatest single source of inefficiency in education" (p.157). He further added, "The number of questions correctly answered on a test is quantitative...but it does not yield a measure appropriate to the behavior of the individual" (p.172).

Conventional testing procedures have been used continuously since the Army Alpha and Beta tests. These were first used in 1917 with U.S. Army recruits for general routine and nonlanguage-scale testing. A conventional test, such as the ACT or SAT, consists of a universe of items to which an examinee is to respond. This universe contains a range of items varying primarily in difficulty.

When a person has been placed into a conventional testing situation, he/she is expected to respond to all items.

It is logical to assume that for nearly every examinee there are items to which he/she is virtually unable to respond. It has been demonstrated by Birnbaum (1968) that the best information about ability estimates comes from those items which the examinee has probability of success equal to 0.5. Therefore, what value or function do these items have in assessing ability?

Conventional test theory models have been based on the concept of true-score. The primary shortcoming has been that the items' parameters change with the norming group. In effect this says that an examinee could take parallel forms of the same test and receive two scores whose difference is statistically significant because the statistics are group oriented. Latent trait theory (Lord & Novick, 1968), which attempts to explain the individual's response to an item, has provided an alternative to conventional test theory.

There are a variety of latent trait models which primarily differ in the number of parameters used to explain individual ability. A latent trait model specifies a relationship between overall examinee test performance and the latent trait assumed to underlie the performance on the test. A latent trait model is a mathematical model since the "unobservable" quantities are described by a mathematical function. The mathematical function is called an item characteristic curve. As a function it relates the probability of success on an item to the ability measured by the item set or the test it contains. It can be demonstrated that the shape of the item characteristic curve (ICC) will

be invariant across randomly selected samples of examinees from a given population.

The one model which attempts to explain a response to an item by an individual in the simplest form and by the most elegant design is the Rasch model. It depends on only two parameters - person ability and item difficulty. The person's ability, the underlying latent trait, is set against the item's inherent difficulty. The functional value is the probability of success of answering the item correctly. This one-parameter model succinctly describes the latent behavior of an individual when he/she responds to an item in terms of probability, and treats the testing situation as a stochastic process. The items are assumed to be stochastically independent. This signifies that the answer to one item will have no information or provide no aid to any other item.

In 1960 Rasch described his model for ascertaining a person's ability when responding to an item on a test (Rasch, 1960). He sought to eliminate two elements which proved to be serious shortcomings for conventional testing. These were person-dependent item difficulty calibration and item-dependent person ability measurement.

In an attempt to alleviate these problems, Rasch designed the concept of "specific objectivity." This concept summarizies the notion of stochastic independence and the notions of instrument-free person measurement and person-free item calibration (Wright,1977). The Rasch model has the following given assumptions about specific

objectivity:

(1) item homogeneity, or unidimensionality,

(2) item characteristic curve,

(3) local stochastic independence.

If these assumptions hold, then the following elements are obtained;

(1) the raw score is a sufficient statistic for ability estimation,

(2) the comparison of two people is fully described by the difference in their abilities on the latent ability dimension,

(3) the estimation of the item difficulties is independent of the ability of the sample on which they were calibrated.

The mathematical formulation of the Rasch model for the logistic ogive curve allows for the estimation of the person ability and item difficulty parameters independent of one another (Rasch 1961,1967; Andersen, 1973, 1977; Barndorff-Nielsen, 1978). When the estimators for person ability and item difficulty are derived by maximizing a conditional likelihood they are unbiased, consistent,efficient and sufficient (Andersen, 1970, 1971). Thus, the parameter estimation results are independent of the sample of persons and the particular items used (Loevinger, 1965).

Traditional test scores allow for prediction of performance only at the point represented by the obtained test score. Properly developed Rasch-based scores allow the user to predict performance on any test item given its

interval scaled distance from the individual's ability score. This principle provides a powerful foundation for the development and use of interpretation procedures that can accentuate the implications to be drawn from educational test scores.

When the central assumption of latent trait theory, that a person's response to an item is determined by nothing other than the person's ability and the item's parametric connection to the latent trait, is accepted by educators, the consequences can be pronounced. Some of these consequences are item banking, test equating, and adaptive testing.

Clearly, the last two decades have been concerned with the notion of adaptive - tailored - testing. This potential for improved measurement was viewed with excitement, presenting a remarkably effective application of latent trait theory. It has been brought about through recent theoretical developments and through steady advances in computer technology. Tailored testing is the fruition of the prediction implied by Green (1970) when he referred to: "the inevitable computer conquest of testing" (p. 73).

Tailored testing is implemented primarily as a computer-interactive method that proceeds as follows: The examinee is seated at a cathode ray tube (CRT) terminal. A test item appears on the CRT and the examinee indicates his/her response by way of the keyboard. If the answer is correct, the next item will be more difficult. If the answer is incorrect, an easier item will follow. With each

response, the computer makes a revised estimate of an examinee's ability. Each revised estimate then becomes a more reliable one.

As mentioned earlier, a major problem, arising because the same test items are given to all examinees, has been that many examinees receive items that are too easy or too hard. If a test item is too difficult, an examinee may resort to random guessing or omission of the response to the item. If the item is too easy, the information available in the response is no more functional than an omitted response. This results in scores containing more error at the extremes of the ability range, indicating that the test is most accurate for average ability examinees. This restricted range of accuracy is reflected in lowered test reliabilities. Thus, tailored testing procedures have been designed to minimize errors of measurement by administering to each examinee only items of appropriate difficulty. This is accomplished by selecting for administration items that maximize the information about an examinee's estimated ability level.

Several methods and procedures have been developed for selecting the items in adaptive testing. Maximum information, stratified adaptive, Bayesian and "up-down" algorithms have been used with varying degrees of success. They all share one or both of the following significant problems for purposes of application in the classroom environment. First, some of the algorithms are dependent on what is termed "prior information." Some previously

determined parameters about an examinee's ability must be known so that the entry-level items can be ascertained. This is difficult and cumbersome for classroom use. Second, the algorithm's computer software is mainframe dependent. Without a computer system with large memory capacities and considerable processing speed, the item selection process and response analyses are very difficult and time consuming.

What is needed to make adaptive testing procedures practical is a microcomputer dependent procedure which can be administered by a classroom teacher. The Binary Search Method (BSM) has been developed with these requirements in mind. The Binary Search Method (BSM) is a procedure for selecting items based simply on the responses to initial items. A full explanation of the BSM algorithm is included in Chapter Three.

Although there have been several documented applications of computerized adaptive testing procedures, very little has been done on microcomputer based procedures or when no prior information on examinees is available. No known research has been conducted using the BSM procedure.

Purpose of the Study

The purpose of the research was to determine if the Binary Search Method (BSM) of computerized adaptive testing could accurately (i.e., reliably and validly) estimate an examinee's ability, to study the effects on attitudes of examinees towards computerized adaptive testing and

conventional paper-pencil testing procedures, and to study the relationship between item response time and item difficulty of the observations provided by the various Binary Search Method selected items. The research was motivated in part by the problems related to the rapidly emerging computer technology, where development and implementation has outstripped research and evaluation.

Additional incentive to study the need for individualized testing has been demonstrated time and again by the biased aspect inherent in group testing procedures. Simulation and field studies have shown reductions in test length and improvements in classification accuracy for computerized adaptive testing procedures. A union formed by the increasingly sophisticated microcomputer technology and adaptive testing procedures based on latent trait theory can provide some of the necessities of individualized testing.

This research was also motivated by computer hardware necessities and sufficiencies such as the need for rapid and accurate computations, clear presentation of the items on the cathode ray tube (CRT) and accurate record keeping of pertinent testing and item response information for examinees.

Questions to be Answered

Through the review of the literature, five research questions were identified that were either supported or rejected by the study:

1. Will the examinee's ability estimate made by the BSM procedure be the same as the conventional test estimate?
2. Will the BSM's estimate of ability be as reliable as the conventional test's estimate, given that the computerized adaptive testing procedure uses only fifteen to twenty percent of the total item set to the make the estimate?
3. Will the examinee's attitude towards testing be significantly different after exposure to the computerized adaptive testing procedure?
4. Does knowledge of the examinee's computerized adaptive testing ability estimate, item response time, sex and previous mathematics grades improve the capability to predict the conventional tesing ability estimate over no knowledge of theses variables?
5. How will the attitudes differ among the examinee's before, during and after the two testing procedures?

Significance of the Study

The research was significant because it involved the viability of the Binary Search Method as a tool for determining an accurate estimate of the examinee's ability with the fewest possible items. Since there was no known evidence surrounding the effectiveness of the Binary Search Method, it was essential that the research base be expanded since recent advances in microcomputer technology have

revived interest in the uses of computers in schools. With the continually decreasing costs and increasing capabilities of microcomputers, the advent of a microcomputer based adaptive testing system was at hand. The need for a classroom oriented computerized adaptive testing system had been felt for many years, but there existed very little in the way of a sound research base for its practical implementation.

Of concurrent interest to computerized testing, when using the BSM procedure for item selection, was the relative efficiency of the items being selected. The relative efficiency is defined as the product of the success probability and its complementary probability (i.e., the probability of failure). This product is used to judge the value of any particular item for measuring a person's ability. This product is transformed into a percentage by multiplying by 400, and the closer this percentage is to perfection (i.e., 100%) the better estimate the item is of the person's ability(Wright and Stone, 1979).

A concern about prior information of an examinee's ability was an additional issue. The source of the prior information has been educational standardized test scores and relevant subject grades taken from examinee school records. The prior information statistics have been used to "prime" the item selection and sequencing procedures of various computerized adaptive testing algorithms.

It is imperative that a strong research base be available to teachers and other educators involved in

testing about computerized adaptive testing applications. This research base must address general questions of efficacy of the procedure and provide a basis of experience for particular implementation models and populations.

Research Hypotheses

Data from student test scores on the ability assessment by the BSM procedure, the assessment of ability by the Pontiac Schools Fifth Grade Mathematics Assessment test and the attitude surveys were used to test the following research hypotheses.

There will be:

1. a significant difference in the examinee's ability estimates when determined by the BSM and the conventional testing procedures.
2. a significant difference in the ability estimates made by the conventional test procedure and the BSM procedure which uses only fifteen to twenty percent of the total item set.
3. a significant difference among the examinee's attitudes as assessed before, during and after the BSM and conventional testing procedures.
4. a significant relationship between the conventional test ability estimate and the BSM ability estimate, the examinees' previous mathematics grades, their sex, and

the examinees' total item response time.

5. a significant relationship among the examinees' total item response time and the BSM ability estimate, the composite experimental test attitude, sex, hardest correctly answered BSM item response time, and their previous mathematics grades.

Limitations of the Study

While this study endeavored to make the BSM algorithm as valid and generalizable as possible, it nevertheless had certain limitations.

The use of the BASIC (Beginners All Purpose Symbolic Instructional Code) language has its limitations in its logical elegance and program structuring. The use of a language such as PASCAL would have probably improved the sequencing speed and data storage of the program. Also, a machine language based algorithm would have provide considerable speed in presenting items to the examinee at the CRT, and it would have also provided much quicker access to and from the disk drive.

The Commodore 64 was limited in its computing power when compared to some of the larger microcomputer systems which have hard disk capabilities. The "64" was fairly suited to individualized testing, but presented some problems when implemented on a large scale basis with many users accessing the system.

Each "64" was equipped with its own disk drive and CRT.

While this had some advantages, data manipulation and collation was hindered by the lack of a more centralized system whereby several computers could have been interfaced with a single hard/soft disk system.

There were only forty-four items on the test. This comprised the entire bank of items for selection and sequencing by the algorithm. Many other adaptive testing strategies have had item banks of one hundred items and more.

The proximity of the two testing procedures may have been a limitation because of the effects of learning on the subsequent procedure. This was minimized by the demand that the post-test be given no sooner than two weeks after the pre-test was completed. The BSM procedure was the pre-test for one group of students and the post-test for another independent group of students of equal size and grade level. Even with this requirement, the results of the post-test in second group were seriously suspect.

The BSM algorithm depends solely on the item's Rasch difficutly estimate for selecting an item. Some authors have suggested that item discrimination, prior information, and other variables should also be used to select subsequent items.

Definition of Terms

Adaptive (Tailored) Testing - the essential idea of adaptive

testing is that the subtest (item) to which a designated examinee is asked next to respond should be chosen in the light of information provided by his or her responses to previous subtests (items).

Analysis of Fit - the study of the plausibility of responses (fit) made by an examinee to a set of items. A t-statistic is calculated on the correct and incorrect responses to determine if there is a significant misfit of the examinee's item response record and score.

Information Function - an index of the relative information with which an observation provides information about the person and item interaction. It is often treated as a percentage of the maximum information that one observation could provide. The information in an observation can be used to judge the value of any particular item for measuring a person. The function has the form: $I(B_v,D_i) = P_i(B_v)(1 - P_i(B_v))$. The range of this function is $0 \leq I(B_v,D_i) \leq 0.25$. See the Rasch definitions below for notations.

Item Characteristic Curve - this curve, which is the graph of the characteristic function, shows the way item i elicits responses from persons of every ability. The curve is invariant across any sample of persons. Its ordinate values consist of the probabilities of a correct response when person v pits his/her ability, B_v, against item i's difficulty, D_i. The domain of definition of both B_v and D_i is negative to postive infinity, as is their difference. Figure 1.1 illustrates an item characteristic curve.

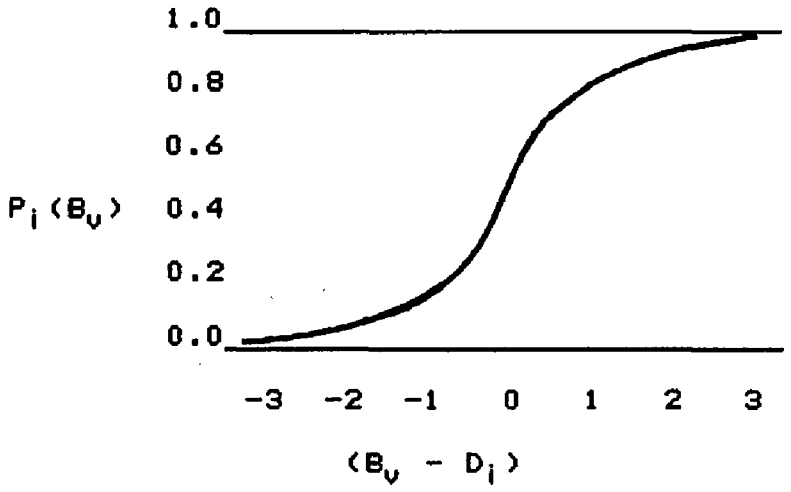


Figure 1.1. Rasch Item Characteristic Curve.

<u>Item Response Theory</u> - this theory is variously designated as latent trait theory and item characteristic curve theory. A fundamental feature of this approach is that item performance is related to the estimated amount of the examinee's "latent trait."

<u>Latent Trait Models</u> - the concept of latent trait is employed in deriving an index of item difficulty, the basic measure being the probability that a person of specified ability succeeds on an item of specified difficulty. The person ability (trait) is a parameter specific to that person and is unobservable (latent). There is no implication that such latent traits exist in any physical or physiological sense, nor that they cause behavior. The latent traits are statistical constructs that arise from a particular analysis of empirically observed relations among test responses. A crude estimate of an examinee's latent trait is the total score he or she obtains on the test. There are other more refined and sophisticated techniques,

such as adaptive testing procedures, that ascertain a measure of a person's latent trait.

Rasch Ability Parameter - an estimate of a person's latent trait, B_v, employing Rasch's theoretical model. These estimates remain invariant despite the particular items used to make the estimate.

Rasch Difficuly Parameter - an estimate of an item's difficulty value, D_i, employing the Rasch model. These estimates, because of their connection with item characteristic curve theory, remain invariant across any particular sample of persons.

Rasch Measurement Scale - the scale is interval-valued. The ability and difficulty estimates are interval-valued as opposed to the conventional statistics of raw score and p-scores, respectively, which are ordinal scaled. When, for instance, two items i and j have the difficulty relationship of $D_i > D_j$, Rasch measurement specifies by how much the difficulties differ.

The Rasch Model - this one parameter logistic model specifies an item characteristic curve which defines the relationship between an examinee's ability measured by the items of a test and the probability of answering an item correctly. This item characteristic curve is represented by the following mathematical expression:

$$P_i(B_v) = \exp(B_v - D_i)/(1 + \exp(B_v - D_i))$$

where: $P_i(B_v)$ = the probability of an examinee with

ability B_v answering item i correctly,

D_i = the item difficulty parameter.

<u>Relative Efficiency Function</u> - An index of the efficiency that an observation, B_v-D_i, can provide about person and item interaction. The function's values range from 0 to 100, and they represent the maximum amount of information provided by an observation as a percentage value. The index is scaled by a factor of 400. This factor is used as a linear transform on the information function's values.

CHAPTER II

A REVIEW OF THE LITERATURE

Introduction

This review of the literature in the area of computerized adaptive testing is concerned with several aspects. These are; 1) latent trait theory and adaptive testing, 2) adaptive testing and computers, 3) various adaptive testing procedures, 4) the issue of item response time and, 5) use of the Binary Search Method for item sequencing and selection.

Latent Trait Theory and Adaptive Testing

Since 1960, there has been a major change in the emphasis of measurement research. Previous to this time research was almost entirely devoted to an elaboration of true score theory and related concepts. However, with the development of latent trait models in the decade before 1960, a major proportion of measurement research and applications has been directed toward latent trait theory. The revolution in computer technology that occurred concurrently has also had an impact on the use of this new

measurement theory.

There has been a strong argument for tailored testing. Ever since the time of the Army Alpha and Beta test, the most common procedure for testing was to gather a group of individuals together to be tested and administer to them a test in paper-and-pencil format. This has been an efficient procedure for testing, but has several inherent problems which need to be addressed. Reckase (1981) has noted several elements of the paper-and-pencil testing procedure which tend to reduce the accuracy of measurement. These elements are broken down into four areas: (a) the artificality of the testing setting, (b) the lack of optimality of the test difficulty for each person, (c) problems in test administration procedures, and (d) test wiseness factors.

The question has become whether or not latent trait theory as proposed by Birnbaum (1968), Lord and Novick (1968), and Rasch (1960) is required for tailored - adaptive - testing. According to Reckase (1981) tailored testing procedures require four methodological components. These include (a) and item pool with item statistics that describe the characteristics of the items; (b) an item selection procedure; (c) a procedure for administering the items; and (d) a procedure for scoring the test that has been administered. Latent trait theories have been used and are well established for (a) and (b), while (c) and (d) have their bases in latent trait theories.

Latent trait theories are not new to educational measurement. Nearly thirty-three years ago Frederic Lord

(1952) introduced the topic of latent trait models to educational measurement specialists. He stated:

> The ability itself is not a directly observable variable; hence its magnitude, in terms of whatever metric may be chosen, can only be inferred from the examinee's responses to the test items. The test score most commonly used as a measure of ability is the sum of the item scores when each response is scored 0 or 1. The metric provided by such a score has the serious disadvantage that it is largely a function of the particular characteristics of the items that happen to compose the particular test administered. (p. 2)

Lord and Novick (1968) went on to say that if an examiner wrote two tests which were comparable in item difficulty and depth and gave them as paper-and-pencil administered tests, chances would have it that two entirely and significantly different scores would be obtained. They stated that "these differences in scores and orderings represent the failure of the measuring instrument to do what we wish it to do."

A theory of latent traits supposes that in testing situations examinee performance on a test can be predicted by defining characteristics of examinees, referred to as traits, estimating scores for examinees on these traits, and using the scores to predict test performance (Lord and Novick, 1968). Since these traits are unobservable and not directly measureable, they are referred to as latent traits or abilities. Cook and Hambleton (1977) stated, "A latent trait model specifies a relationship between observable examinee test performance and the unobservable traits or abilities assumed to underlie performance on the test" (p.

7).

There are several different latent trait models which are in vogue for different types of procedures of adaptive testing. One is the three parameter model first described by Birnbaum (1968). The three parameters are item difficulty, item discrimination, and item guessing. These three are used to determine the probability of success per item by an individual with a given ability. There are certain problems with the three parameter model. One of these was the uncertainty of the parameter estimates since they do not converge at the high and low ends of the item difficulty estimates. Wright (1977) indicated, "All other latent trait models [except the Rasch model] lead to more complex scoring rules that involve unknown parameters for which satisfactory estimators do not exist" (p. 45).

The Rasch model is a single parameter model dependent only on the item difficulty parameter. With the ability parameter "the two parameters represent the positions of persons and items on the latent variable they share" (Wright, 1977, p. 44). They are used in the Rasch model to determine the probability of a person with a given abiltiy to succeed on an item with a given difficulty (Rasch, 1960).

The one and three parameter models, as proposed and solved by Rasch and Birnbaum, respectively, have differing solution procedures for estimating their respective parameters. Several studies have been completed which have compared and contrasted the two models' parameter estimation procedures. Benson (1982) has surveyed the research in this

area and found through his own empirical study that there is no significant difference between the two models' item difficulty parameter estimates when used to determine test efficiency.

The advantage of the Rasch model rests on the notion that "the level of performance on an item is governed by two factors; by the ability of the person taking the item and by the difficulty of the item" (Gusstafsson, 1980). A slightly higher precision of measurement can be achieved through the following statements:

1. For any given item, a person with a higher ability should have a higher probability of passing the item than a person with a lower ability.
2. For persons with the same level of ability, the probability of passing an easier item should be higher than the probability of passing a more difficult item (Gusstafsson, 1980).

Traditional test scores allow for performance prediction only at the point represented by the obtained test score. Properly developed Rasch-based scores allow the user to predict performance on any task given its interval-scaled distance from the individual's ability score. The Rasch model thus has the main ingredients for providing the necessary and reliable item parameter estimates.

Rushby (1979) stated, "Most statistics which are used to describe item performance are only accurate in the

context in which they were obtained. The Rasch method has overcome this difficulty and permits item analysis which is independent of the context in which it is used" (p. 274). Mead (1979) pointed out that the Rasch model is the simplest and most elegant logistic model known and, "It is the very definition of measurement" (p. 172). These concerns about the test context and person-free measurement by a set of items was elaborated extensively by Rasch (1960, 1966). He referred to this particular capacity of his model as "specific objectivity."

If the group to be tested is sufficiently heterogeneous, it is impossible for a conventional test to measure accurately at both high and low ability levels at the same time. To obtain effective measurement at the low or high ability levels, there must exist low or high difficulty estimates on items. When number-right scores or conventional formula scores are used, the hard items not only waste the time of the low ability examinees, they impair whatever measurement of these examinees would otherwise be effected by the easy items. A similar argument holds for high ability examinees and low difficulty items. Weiss (1982) stated, "Conventional tests tend to overestimate low B levels and to underestimate high B levels" (p. 119).

If the desire is to measure accurately throughout a wide range of ability, there is a need to match the difficulty level of the items administered to the ability level of the examinee being tested. This is what is defined as tailored or adaptive testing. As Reckase (1981)

indicated, "The item parameter statistics have to be dealt with sufficiently, and it is believed that the Rasch model's mathematical formulation of the item estimates has accomplished this, and provides the necessary information to be used with an adaptive testing strategy" (p. 74).

Adaptive Testing and Computers

To facilitate the use of adaptive testing, the computer has been implemented as the selector, deliverer and analyzer of the test items. Several important and positive qualities for tailored testing have been contributed by computer implementation. These qualities are test security, item selectivity, individualized testing, testing environment, and dynamic item sequencing and analyses.

Several very useful aspects of the computer have been demonstrated in the testing setting. Sampson's (1983) list of potential benefits that can contribute to better measurement by using the computer is:

1. positive client response,
2. cost effectiveness,
3. uses in a variety of adaptive testing situations,
4. more efficient use of staff personnel,
5. test scoring and administration efficiency,
6. reduced scoring error rates,
7. assistance to individuals with visual, auditory, and physical limitations.

Jelden (1982) also demonstrated several benefits by using the computer in a testing situation. First, he said, "Computer assisted evaluation is an adjunct to a more realistic approach to learning, that of individualized instruction" (p. 19). He further pointed out that there are certain computer assisted evaluation options, such as:

1. individualized cognitive testing,
2. almost endless variations of a test from an item pool,
3. freedom by test designers to add or delete items from an item pool,
4. updated item analysis,
5. generation of sample test.

Along with the potential benefits of cost effectiveness, aid to special students, and better test design mechanisms, there appear to be other factors which make the computer a powerful device for testing. Schmidt, Urry, and Gugel (1978) indicated that there was increased test-taking motivation and increased interest and favorable attitudes toward testing and its method of administration when testing was performed on a computer. Urry (1977) stated, "The use of computer-aided testing can be made to suit individual plateaus for test differences among examinees" (p. 211). He further indicated that after a few preliminary items the computer algorithmically sequences questions to the examinee at his/her level and degree of difficulty.

On a different level and point of view, McBride (1979)

found that traditional test stimuli are static and usually of a single color when presented to the examinee. He also claims, "Spatial perception, short term memory, judgment, integration of complex stimuli, cognitive information processing, and other complex abilities could be measureable by using the computer as a testing medium" (p. 153).

More pertinent to the requirements of testing, such as reliability and validity, several researchers have pointed out research benefits by using the computer in adaptive testing. Hansen (1969) claimed, "[Computerized testing] should expect conventional measures of reliability to improve for at least two reasons: 1) core variances should be greater due to the spread created by branching and, 2) the effect of sequential branching should more optimally match the examinee's performance levels with test item difficulty levels and reduce guessing behavior" (p. 229). He also stated that a low correlation (r=0.21) was observed between computer item answering time and conventional test time. This implies different strategies used by examinees as individual differences when responding to test items. Spineti and Hambleton (1977) noted a reduction of at least 50% in testing time.

Certain aspects of computer testing have been quite appealing from the point of test-taking attitudes and anxiety producing effects of conventional testing. Katz and Dalby (1981) demonstrated in an interesting study that test anxiety was significantly reduced when items were presented to groups of young students by means of the computer. They

also showed that test administration time was significantly reduced.

Test security, including the prevention of access to test copies prior to testing time has been a perennial test problem. However, with the computer testing environment as a proven and viable alternative these difficulties are significantly alleviated. Green, et.al. (1982) noted that there is considerable improvement in test security when testing is performed by computer presentation. With computerized testing, there is no hard copy and no preserved list of items to which the examinee can gain easy access. Again, there has been concern with the cognitive domain of the testing situation. Gren (1970) said, "Computer testing offers a better understanding of cognitive processes by analysis of response sequences and item sequences" (p. 213). Green's statement implied that the compute, as the vehicle for testing, can efficiently and accurately collect examinee response information and correlate it with various diagnostic and prescriptive elements, germaine to further testing or remediation. Edwards (1980) more recently pointed to the fact that there is "An additional advantage of the microcomputer systems that results from the control of the hardware by persons who have expertise in education" (p. 72). This points out that most previous research has used a computer mainframe-type system to present the items. Very little research on adaptive testing with a microcomputer system has been conducted.

McKinley and Reckase (1980) showed that adaptive

testing requires the use of the computer for presentation of test items since adaptive testing is an interactive process. Any paper-and-pencil method of adaptive testing would default to a bureaucratic quagmire before any meaningful results could be realized. They further noted that the computer and its software algorithm handles such viable concerns of testing as entry point into the item pool, the procedure for estimating ability, the item selection method and a rule for terminating the test. All this would necessitate a highly qualified person to singularly administer the test to a single examinee if the computer was not used.

Nicewander, et.al. (1980) reported, "Research has shown that in terms of providing adequate information across nearly all ability levels, tailored testing has no competitors" (p. 175). Reckase (1981) summarizes the necessity of the computer being used in adaptive testing:

> The advantages of such a procedure are potentially very great. First, since the tests are given individually, they can be self-paced and self-scheduled. This removal of time constraints may also help in reducing the stress inherent in the testing environment. Further, by placing the screen and keyboard in a cubical the work environment can be greatly improved while at the same time increasing test security. Test security is also enhanced because each person gets a different test and no paper copies of the test are ever produced. (p. 37)

Different Adaptive Testing Procedures

The computer, as has been demonstrated above, has become the facilitator of adaptive testing. There has been

much research done in this area using several different procedures, algorithms, and methods for selecting items, termination of testing, analyzing item responses and estimating abilities.

At first conventional tests were presented on the computer, but, as has been pointed out by Weiss (1982), presenting a test of fixed item length presents difficulties with items either too difficult or too easy. These items act as confounding effects for analytical purposes and act as anxiety-producing elements during the test which certainly produce errors in measurement.

McKinley and Reckase (1980) described a method of item sequencing and selection which has been termed a maximum information function method. The method presents different items from the median of the item pool where the items have been arranged from least to greatest difficulties as determined by Birnbaum's (1968) three parameter logistic model. The sequencing of the items (whether the examinee is given an easier or a harder item) depends on the response of the individual. The items selected are based on those providing maximum information and the current estimated ability of the examinee. The information function used programmatically is according to that of Birnbaum's three parameter logistic model (1968).

Urry (1977) described a method of item selection and sequencing based on the criterion of reducing the standard error of the estimate as the items seek that postion of difficulty which is nearly equal to that of the estimated

ability of the examinee.

Waters (1977) utilized a procedure which he called the "stradaptive" - stratified adaptive-method. This item sequencing procedure created strata of item mini-pools each of which had narrow-interval item difficulty levels. He tested three termination rules with differential termination functions. The results culminated in a significant reduction in the number of items needed to obtain reliable and valid measures when compared to a conventional test.

The Bayesian sequential procedure, as first proposed by Owen (1975), selects items based on a Bayesian prior information function and conditional prior ability information about the examinee. However, Thompson and Weiss (1980) demonstrated that the stradaptive method is a more valid ability estimator than the Bayesian procedure. Also, McKinley and Reckase (1981) showed evidence that the Bayesian procedure of tailored testing was biased to the mean of the prior distribution. In addition, they noted that false assumptions of prior ability estimates significantly altered the adaptive ability estimates. These prior ability estimates were used in both the maximum information method and the Bayesian method.

Lord (1970) described a method of stochastic approximation attributed to Robbins and Monro (1951). The method administers an item from the median of the difficulty ordering, giving a harder item for a correct answer and an easier item for an incorrect answer. At the beginning of the procedure, the difficulty degree changes are large, but as

the testing proceeds, these differences in the difficulty of the successive items go to zero as the items whose difficulty level are such that the examinee would successfully answer those items 50% of the time.

Cudek, McCormick, and Cliff (1980) utilized an implied orders method of item selection. Briefly, it states that a person A, who answers item B correctly, and a person C, who answers item B incorrectly, have the following relationship: A > B > C. This says that person C should attempt as his/her next item one which is median to B and C, while person A should attempt an item which is median to A and B. They noted that "95% of all responses implied by the tailored procedure were identical to responses actually observed."

Sympson and Weiss (1982) made use of the microcomputer as the hardware medium of a maximum information function strategy. Their's was the only investigation known that made use of a microcomputer system. All others depended on the mainframe-type system with its tremendous computing power for various analyses and item pooling. They demonstrated that the fixed-entry point of the adaptive test must assume an ability estimate of B = 0.0 (i.e., mean ability parameter of a normal distribution), assuming further that a person has been drawn from a normally distributed population, and no additional information was available about a person's ability on the given test dimension. Variable-entry point assumed prior knowledge of a person's ability which had been presumably obtained on an independent measure. Pertaining to test length and measurement precision, it was found that an

adaptive test was only one-third to one-half the length of the conventional test, and the adaptive method they employed was a reliable estimator of examinee ability, similar to the conventional test. Finally, their study showed that the adaptive test provided criterion-related validity equal to the conventional test.

Rosso and Reckase (1981) studied the length of an adaptive test and its reasonable estimate of ability of the examinee. They showed that one could obtain reliable ability estimates with the use of 12 to 14 items algorithmically selected from a pool of some 150 items, again ranked by their difficulty parameter estimates using the one- or three-parameter model.

While these and other adaptive selection strategies have been employed, Wetzel and McBride (1983) concluded through extensive research, "Item selection strategies that explicitly imply optimization criteria should be regarded as preferrable to simpler strategies that do not" (p. 83).

Item Answering Time

Of crucial concern has been the issue of item answering time. Some item response time research has been conducted both by Monte Carlo simulation and under field experimental conditions. While a person taking a conventional test has a fixed time to answer all items, the adaptive examinee is not restricted by overall time limitations. The objective of Water's research was to see if there was any relationship

between the correct/incorrect responses and the length of time it took for the examinee to respond to the item. Waters (1977) recommended in his research that testing time, rather than number of items, be used as a dependent variable for adaptive testing research. With the capabilities of the microcomputer this assessment is a direct and easily measured variable. McKinley and Reckase (1981) showed that response time was greater for incorrect responses than for correct ones. They recommended that this issue be researched for its relationship to test-taking strategies that are employed by different examinees.

No known research has been done in the area of discrimination values on items, incorrect response selections with multiple choice items, or the response time factors.

Use of the Binary Search Method

The Binary Search Method (BSM) was based in part on the maximum information function described above. The items are ordered by their Rasch item difficulty parameter estimates according to increasing difficulty. A median difficulty item is selected and the student is given a harder item following a correct response, or an easier one following an incorrect response. The objective is to seek out the item that most closely approximates the ability of the examinee. Statistical theory shows that maximum variance is taken into account when the probability of a correct response is 0.5.

This probability occurs exactly when the ability of the examinee is equal to the difficulty estimate of the highest correctly answered item by the examinee.

The information function, $I(B) = P(B)(1-P(B))$, has its maximum value when the probability $P(B) = 0.5$. This function governs the selection process in the BSM. Thus, the first item presented to the examinee is a median item from the item difficulty sequence. Spineti and Hambleton (1977, p. 153) showed, "It is best to start a[n adaptive] test at the middle of the learning hierarchy," when the item difficulty parameters have been arranged on a "hierarchial" basis according to the Rasch difficulty estimates.

Summary

As has been stated above, there are several different strategies for sequencing and selecting items from a larger or smaller item pool. The successful strategies, those with high reliability when compared to the conventional test, all employed optimal selection and sequencing strategies and, most notably, the use of a large main-frame computer system. Unfortunately, this is a cumbersome way to employ computerized adaptive testing strategies.

Testing, by and large, takes place in a school classroom where microcomputers are ideally suited both for cost effectiveness purposes and for teacher supervision of the testing milieu. Dependency on a mainframe-type system is difficult to operate and connection/dependency problems

cause difficulties for the average classroom teacher. This could result in a decline in the use of the adaptive testing situation, regardless of how sophisticated the strategy. The teacher would probably return to the conventional testing paradigm thereby negating all the efforts made for implementation of adaptive testing.

The choice of which latent trait theory to use appears to be largely a matter of personal opinion based upon some limited factual basis rather than scientifically firm testing theory. The Rasch model was used as the means for item difficulty parameter estimates since there is no significant difference between estimates made by this model and the three-parameter model. Also, the Rasch model has the essential criteria of specific objectivity and the data that were used for this study fit the model's criteria very well.

CHAPTER III

RESEARCH METHODOLOGY AND DESIGN

Introduction

This study compared the estimates of an examinee's ability level - latent trait - as determined by the Binary Search Method(BSM) of computerized adaptive testing, and the actual ability estimate determined by the conventional paper-and-pencil test which contains the "universe" of items.

The means of testing the efficacy of the BSM procedure was to check the results of this procedure against the examinee's score on the entire test taken in paper-and-pencil form. If the score or ability level estimate determined by the BSM was not significantly different from the same examinee's score on the paper-and-pencil test, then it was inferred that the BSM computerized adaptive testing procedure yields a valid and reliable measure of the estimate of the examinee's latent trait.

If this last conjecture held and accurate estimates of an examinee's ability on the test's single dimension were obtained a person's latent trait on this single dimension could be accurately estimated by fifteen to twenty percent

of the entire universe of items.

This study also addressed the issues of test item response time, and maximum item information. The question of response time is a function of several variables, such as BSM ability estimate, conventional test ability estimate, and others. Maximum information in an item is determined by Birnbaum's (1968) formula for the one-parameter model. The conventional test ability estimate was used as the "true" examinee score and the highest correct item on the BSM procedure was used in the calculation of relative information. Of equal importance, examinee attitudes toward testing will be surveyed before, during, and after the testing sequence.

Description of the Study

This research study was experimental in nature. It was conducted to determine if a given method of testing, namely the Binary Search Method of computerized adaptive testing, could accurately estimate an examinee's latent trait on the single dimension of the test by sequencing the items within the neighborhood of the true latent trait parameter of the examinee. The single dimension of the test is defined by its items. The neighborhood of the latent trait parameter is the confidence interval estimate of the latent trait parameter as determined by the Rasch model estimation procedure (Wright, Mead, & Bell, 1979).

A paper-and-pencil test was utilized to obtain a valid

estimate of the examinee's ability. The paper-and-pencil test used in this study was the Pontiac Schools Student Assessment Program - Mathematics Test: Grade 5. This test was be administered to all students who participated in the study. This test contained a total of forty-four items each with a different Rasch difficulty estimate. The range of the item difficulties was -1.8 to 1.9. This set of items was the "universe" of items for the experimental procedure.

The microcomputer system utilized in the study was the Commodore-64 with a CRT and a Commodore disk drive (model 1541). The software employed in the computerized adaptive testing on the microcomputer system was the Binary Search Method (BSM). This software, which selected and sequenced the items for CRT presentation, collected and recorded all pertinent data, and contained the test termination rules, was written by the author.

The BSM item selection and sequencing procedure is a simple but elegant binary search of the items, which have been ranked according to the items' interval-valued Rasch model difficulty estimates.

As with any binary search technique, the BSM procedure was based on the initial idea that from a given point of origin there were two ways to proceed pursuant to a reaction (i.e., response). This "binary" approach continues until a desired result has occurred. The BSM procedure begins the testing of an examinee at the "middle" of the ordered set of items. The ordering was based on the Rasch difficulty estimate of each item. The initial item was chosen from one

of two items whose Rasch estimate was 0.00 or 0.03 on the Rasch measurement scale. These two initial items were of median difficulty, and were used assuming a normally distributed population. The second item selected for presentation to the examinee was a function of the examinee's response to the initial item. If the response was correct, $X_{vi} = 1$, for person score v and item i, then the next item was half way between the hardest item and the last item correctly answered. If the response to the second item presented was incorrect, $X_{vi} = 0$, then the third item for examinee response was halfway between this last incorrectly answered item and the last correctly answered item. This binary search for items continued until the last correct item is one less in order value then the lowest incorrect item. Accordingly, this establishes the termination criteria. Figure 3.1 lists two illustrations of the item sequencing.

Illustration 1.

Attempted items[a]: 21, 12,16,14,15

Response values[b]: 1, 2, 1, 2, 2
(HCI)

Graphics on item sequencing: top arrows for correct response, bottom for incorrect.

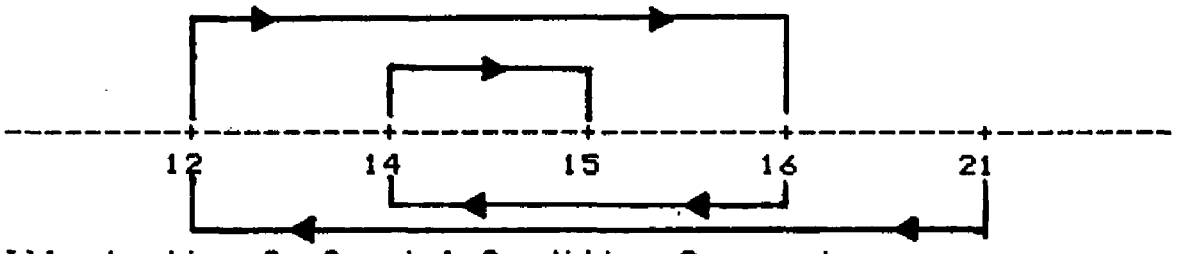
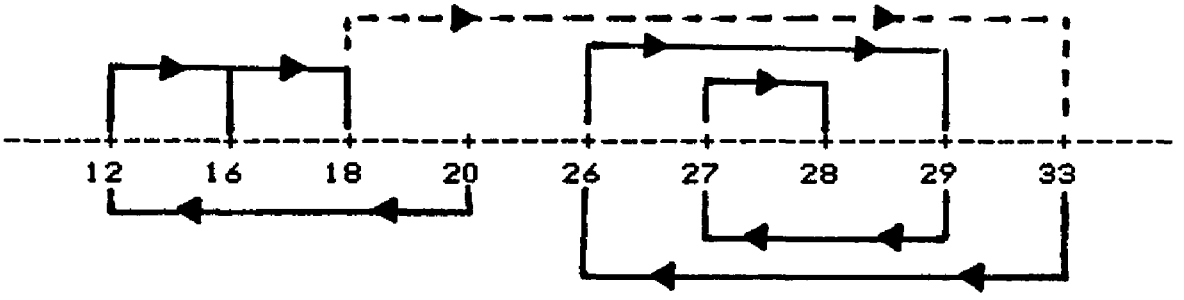


Illustration 2: Special Condition Sequencing.

Attempted items: 20, 12, 16, 18, 33, 26, 29, 27, 28

Response values: 1, 2, 2, 2, 5[c], 6, 5, 6, 6
(HCI)
(this illustrates a false negative initial response)



Note: a - item presentation sequence to the student. The numbers are the BSM enumeration of the Rasch item difficulties in ascending order
b - answered correctly if even value, incorrectly if odd
c - coding value for special condition sequence

Figure 3.1. Diagrams illustrating the BSM algorithm for normal and extended item sequencing.

Instrumentation

The instruments used for measurement in this study were the Pontiac Schools Student Needs Assessment Program -

Mathematics Test: Grade 5, hereafter referred to as the conventional test (or SNAP5), the Binary Search Method of Computerized Adaptive Testing (an author written and developed computer algorithm for item sequencing, selection, and grading), hereafter referred to as BSM, and the Hungerman (1965) Mathematics Attitude Survey, hereafter referred to as MAS. The statistical analyses and interpretation of results from these intruments and their practical significance in relation to the problems being investigated was an integral part of this study.

The conventional test, SNAP5, is a forty-four item instrument designed by and for Pontiac Schools to suit the objectives of the fifth grade mathematics curriculum. After testing was done with all fifth grade students in the fall of each academic year, the items were subjected to analysis for difficulty values, person ability parameter estimates, and other pertinent item analyses, using the Rasch model. Each item has an interval valued difficulty parameter estimate. From these estimates an ability score is rendered for each student.

The BSM testing procedure made use of the item difficulty parameter estimates to select and sequence the items to the examinee. The forty-four items were rearranged for the BSM procedure by their increasing Rasch difficulty values.The correctly answered item with the highest difficulty was the person ability estimate for that examinee. Two special conditions, Condition I and Condition II, were integrated into the item selection and sequencing

procedure. Condition I dealt with a false negative initial response, while Condition II dealt with a false positive initial response. A false negative response to the initial item by the examinee meant that he/she would have responded correctly a majority of the time. The erroneous response was probably due to several factors, such as typing the wrong key, misreading the choices, and so on. A corresponding definition was given for the false positive response. These two conditions, I and II, were built into the BSM algorithm to control for these potential problems.

The Rasch difficulty values are interval-scaled. These difficulty estimates differ from the classical p-scores. The p-scores indicate that one item is harder than another, but the Rasch values indicate by how much.

The Hungerman MAS (Hungerman, 1965) was adapted for fifth-grade level from the Aiken-Dreger, Likert-type Mathematics Attitude Scale (Aiken and Dreger, 1961). This revised instrument was subjected to a test-retest study in two fifth-grade classes in May, 1984 with a reliability of r = 0.93. It posed no reading problems and was completed in a relatively short time by all students. The students were instructed that they could take as much time as necessary to complete the survey. The MAS consisted mainly of twenty items, ten positive and ten negative, testing attitude towards mathematics. There were also five items at the end which obtained a single response each, for feelings regarding reading, social studies, science, arithmetic, and "all subjects." In addition, five items on computer

attitudes were written and added by the author. Answering the twenty basic mathematics items with the most agreeable reaction value of 5 resulted in a maximum total score of 100. The least favorable reaction to twenty items would result in a minimum score of 20. The five computer attitude items were summed for attitude tendencies toward computers and there relationship with other variables.

Design of the Study

The design of this study consisted of two groups of fifth-grade students who were assessed using three instruments; SNAP5, BSM, and MAS. The two groups, A and B, were tested by the BSM procedure and the SNAP procedure. Group A was given the BSM procedure first, followed by the SNAP5 procedure. The sequence of the two procedures was reversed for Group B. There was a minimum of three weeks between procedures for both groups. The size of each group was virtually the same. Figure 3.2 illustrates the design paradigm for the ability assessment part of the study.

Procedure		Group A	Group B
	BSM	Pre x	
		Post	x
	SNAP	Pre	x
		Post x	

Figure 3.2. The design for test data collection of the two procedures for both groups.

The attitude assessment was made for both groups A and B. The assessment was made before the pre-test, between the pre-test and post-test, and after the post-test. Groups A and B were assessed with the MAS attitude survey instrument before, during and after the administration of each of the two testing procedures. The design paradigm for the attitude

	Before	During	After
A	x	x	x
B	x	x	x

Figure 3.3. The design for collection of MAS attitude data.

Several variables were identified for use in this study. They were used to analyze the two testing procedures and the attitudes of the examinees. The variables are listed

below:

BSM - ability estimates made by computerized BSM algorithm procedure.

SNAP5 - ability estimate made by the conventional test SNAP.

ATT1, ATT2, ATT3 - attitude assessments of groups A and B made before, between, and after the pre-test and post-test procedures,respectively.

GRADE - number of mastered objectives by the examinee on the mathematics section of the 1983 Michigan Educational Assessment Program.

IRT - mean item response time for each examinee's item vector on BSM procedure.

IRTMAX - maximum item response time for all items in the BSM item vector.

DIFFMAX - Rasch item difficulty estimate of the highest correctly answered item on the BSM procedure.

INFOMAX - maximum value of INFO(I).

SEX - male or female

Systematic Procedures

The study was conducted during the first semester of the 1984-85 adademic school year in twenty fifth grade classes in eight elementary schools of the Pontiac School System.

The conventional test, the Pontiac Schools Student

Assessment Program - Mathematics Test: Grade 5, (SNAP5), consisted of forty-four items, each with a Rasch valued difficulty parameter estimate. Rasch-based ability estimates were determined and were equated with each raw score, except a perfect right score and a perfect wrong score. Perfect right and wrong scores, those which were all right or all wrong, respectively, provide no information about the examinee on the defined ability dimension of the test. The conventional test item numbering sequence was not associated with the low-to-high ordering of the Rasch estimates. For the BSM procedure, the items are numbered by the low-to-high ordering sequence. It was this ordering that was used for the BSM item selection algorithm.

The classes were randomly chosen to include at least four hundred fifth grade students. Each class was provided with a computer and all necessary peripheral hardware. There were certain questions on the conventional test which, when presented on the CRT, needed external devices for clarity. These devices consisted of printed material containing figures, drawings and other test-pertinent elements needed by the examinee for clarification of the particular test item.

For the experiment, approximately one-half of the students involved were given the BSM procedure first. The conventional testing procedure was then administered to all students. A second group was tested in reverse order. The time interval between procedures was no less than two weeks and no more than five weeks. These time restrictions are

necessitated by the possible effects of testing on the conventional procedure, and to reduce the interference the experiment would have on normal classroom operation. This necessitated starting the BSM procedure five weeks prior to the fixed date of the group testing by the conventional procedure. The BSM testing procedure was completed two weeks prior to the conventional testing date. The reverse testing sequence for the second group was conducted under the same time constraints.

During the administration of the BSM procedure, several data elements were collected for each examinee which were used for various analyses. These were response time for each item, an item sequence/response matrix, total test time, and a coding of responses if special conditions of item sequencing are encountered. Data for each examinee was stored by the microcomputer system's disk storage device.

Prior to the entire testing sequence, students were given an attitude survey concerned with their personal feelings about taking tests. It was assumed that none of the experiment's participants had ever been involved in any computerized testing situation prior to this rearch study. Between the two testing procedures and after the entire testing sequence, students were asked their attitudes on being tested by conventional means and by computer means. A series of semantic differential scales were prepared for this purpose.

Teachers were given as much flexibility as possible in implementing the experiment. In order to facilitate the BSM

procedure for the students, the teachers gave the students all the assistance necessary for proper entry of responses. This assistance was provided to remove the confounding effect of an examinee's ability to operate the equipment. It was not part of this study to ascertain this ability. During the experiment teachers of all classes were asked to conduct class as usual to avoid any confounding effects on the testing procedures. Students were placed at the computer for the BSM procedure by the teacher to ensure a proper testing environment. Due to equipment limitations, no more than two students were tested at any time. Teachers were asked to note any hardware or storage problems during the procedure. Students were told that the computerized adaptive testing procedure was being used only for verification of the conventional procedure. Further, teachers were asked to note any difficulties students might have with the computerized adaptive testing procedures. Complete logs of students activities at the computer were kept by the computer and the teacher.

Pilot Studies

A pilot study of the BSM procedure was made in May, 1984. Two classes of fifth-grade students were utilized for the purpose of studying the algorithm's performance before actual data collection was attempted. An analysis of the results of this pilot study demonstrated the sufficiency of the BSM method. At the same time, the adapted version of the

MAS was given to all pilot study students for the purpose of collecting data to compute attitude/score correlations.

A Monte Carlo simulation of 2000 "simulees" was made in January, 1984, utilizing the same equipment used for the study. The simulation results were satisfactory for pursuing the procedure for field study.

Sample and Population

The sample to be used for data collection consisted of students in the fifth-grade who attend one of eight elementary schools in the Pontiac School System.

There was a total of eight schools utilized with a combined total of twenty classes. These classes allowed for the assessment of some 300-350 examinees.

The twenty classes were randomly divided into two groups of ten classes each for the different testing sequences. None of the eight schools were divided between the two groups, A and B.

Field Procedure

The BSM procedure necessitated the use of a reliable and cost effective microcomputer system for selecting, sequencing, and presenting the items to the examinee. Also, the system had to collect, analyze, and store the data for each examinee in an efficient and reliable way. This task was made easy by the use of the Commodore 64. Interfaced

with the "64", is a "green-screen" monitor (CRT) and a Commodore Disk Drive, Model 1541.

The Rasch-ordered items, individual student files and all computer programs were contained on a single diskette for each class. While the examinee was seated at the computer terminal, the diskette was activated to present items, and upon completion of the BSM testing procedure, the individual examinee's item response matrix was recorded on the diskette. All programs, including the BSM algorithm, were written, and tested for efficacy through a pilot study and simulation by the author.

So that the contaminating effect of computer-usage competency was minimized, the examinee was presented with an easy introduction to how he/she should enter answers to the individual items. These practice items were constructed by the author and were parallel to items on the actual test. These practice items had Rasch difficulty values of -1.0 ± 0.2.

The examinees were individually tested in an area of the classroom which provided the examinee with a sufficiently calm atmosphere to concentrate on the test items. This aspect was a responsibility of the classroom teacher.

The teacher could provide guidance to the examinee in the form of non-item assistance. The teacher was not allowed to help the student with any item, nor provide any clues to arriving at an answer.

The SNAPS procedure and the MAS assessment were given

to each class on a group basis. The conventional paper-and-pencil method of data collection on both instruments was used. Neither the SNAP5 test nor the MAS were timed instruments, although both were completed by all students within forty-five and fifteen minutes, respectively.

The teachers were asked to keep logs of the daily activities while the examinees were being assessed by the BSM procedure. Log items included: when an examinee was at the computer, unusual conditions arising at any time, and teacher observations and comments.

Questions and Statistical Analysis

The examination of the data consisted of an analysis of covariance on the BSM ability estimate and the conventional ability estimate, intercorrelation of several of the aforementioned variables, a multiple regression on several dependent variables, and t-Tests and Chi-square analysis on attitude assessments before, during and after the two testing procedures.

The research questions are stated below with their accompanying null and alternative hypotheses. The questions, stated initially in Chapter I, are restated in statistical form.

QUESTION 1. Will the examinee's ability estimate made by the BSM procedure be statistically the same as the conventional test (SNAP5) estimate?

H_0:1.1 - There is no significant difference in the examinee's ability estimates when determined by the BSM and SNAP5 ability estimates.

H_1:1.1 - There is a significant difference in the examinee's ability estimates when determined by the BSM and SNAP5 ability estimates.

H_0:1.2 - There is no significant relationship between the BSM and SNAP5 abilitly estimates.

H_1:1.2 - There is a significant relationship between the BSM and SNAP5 ability estimates.

<u>QUESTION 2.</u> Will the BSM's estimate of ability be as reliable as the SNAP5 estimate, given that the computerized adaptive testing procedure uses only fifteen to twenty percent of the total item set to make the estimate?

H_0:2.1 - The use of fifteen to twenty percent of the SNAP5 items will not give a significantly reliable estimate of the examinee's ability.

H_1:2.1 - The use of fifteen to twenty percent of the SNAP5 items will give a significantly reliable estimate of the examinee's ability.

<u>QUESTION 3.</u> Will the examinee's attitude towards testing be significantly different after exposure to the computerized adaptive testing procedure?

H_0:3.1 - There is no significant difference in attitude towards testing in mathematics after the examinee has been exposed to computerized adaptive testing.

H_1:3.1 - There is a significant difference in attitude towards testing in mathematics after the examinee has

been exposed to computerized adaptive testing.

QUESTION 4. Does knowledge of the examinee's BSM ability estimate, item response time, sex and previous mathematics grades significantly improve the capability to predict the SNAP5 ability estimate over no knowledge of these variables?

H_0:4.1 - The total item response time is not significantly related to the examinee's BSM ability estimate.

H_1:4.1 - The total item response time is significantly related to the examinee's BSM ability estimate.

H_0:4.2 - The total item response time is not significantly related to the examinee's SNAP5 ability estimate.

H_1:4.2 - The total item response time is significantly related to the examinee's SNAP5 ability estimate.

H_0:4.3 - The total item response time is not significantly related to the examinee's previous mathematics grades.

H_1:4.3 - The total item response time is significantly related to the examinee's previous mathematics grades.

H_0:4.4 - There is no significant relationship among an examinee's previous mathematics grades and the BSM ability estimate, the SNAP5 ability estimate, and the initial-item response time.

H_1:4.4 - There is a significant relationship among an examinee's previous mathematics grades and the BSM ability estimate, the SNAP5 ability estimate, and the initial-item response time.

H_0:4.5 - There is no significant relationship among the SNAP5 ability estimate and the BSM ability estimate and the examinee's previous mathematics grades.

H_1:4.5 - There is a significant relationship among the SNA5P ability estimate and the BSM ability estimate and the examinee's previous mathematics grades.

H_0:4.6 - There is no significant relationship between the SNAP5 ability estimate and the highest correctly answered BSM item difficulty value.

H_1:4.6 - There is a significant relationship between the SNAP ability estimate and the highest correctly answered BSM item difficulty value.

H_0:4.7 - There is no significant relationship between the BSM ability estimate and the SNAP5 ability estimate, and the examinee's previous mathematics grades.

H_1:4.7 - There is a significant resationship between the BSm ability estimate and the SNAP5 ability estimate, and the examinee's previous mathematics grades.

<u>QUESTION 5.</u> How will the attitudes differ among the examinees before, during, and after the two testing procedures?

H_0:5.1 - There are no significant attitude differences within the two groups before, during or after the two ability estimation procedures.

H_1:5.1 - There are significant attitude differences within the two groups before, during or after the two

ability estimation procedures.

Methodological Assumptions

The following assumptions have been made about the method used in this study:

1. There was no dependency on previous computer knowledge by the examinee.
2. The ordering of the items by their Rasch difficulty estimates was made based on the theory of the Rasch model and its assumptions and conditions.
3. The sequencing of the items was made by the BSM algorithm which used the Rasch difficulty parameter estimates for determination of the sequencing and selection

CHAPTER IV

FINDINGS

The findings of the study are presented in Chapter IV. Pearson product moment correlations, multiple regression equations using standardized z-score variables, and Student t-tests' values were calculated.

Findings are shown in the same order as the hypotheses. Separate tables display the findings for each hypothesis.

Scattergram plots and their respective correlation information for the two groups involved are included in Appendix B.

The two groups, A and B, are dealt with separately throughout the findings, except where indicated. Group B, those examinees who took the SNAP5 test first, were largely used as a control and contrast group for studying the primary sample, Group A.

Hypothesized Findings

Hypothesis H_0:1.1;

> There is no significant difference in the examinee's ability estimates when determined by the BSM and SNAP estimates.

In order to test this hypothesis, a t-Test was conducted to determine if any statistical significance existed between the ability estimates when made by the BSM and SNAP5 procedures.

The null hypothesis was accepted for Group A, while being rejected for Group B. Table 4.1.1 displays the t-Test results for the two-tailed probability.

For Group A, the mean and standard deviation for the BSM procedure was 95.98 and 5.49, while the mean and standard deviation on the SNAP5 was 96.16 and 6.68. A two-tailed t-Test was not significant at $p<0.05$.

For Group B, the mean and standard deviation for the BSM procedure was 96.48 and 5.08, while the mean and standard deviation on the SNAP5 was 95.75 and 5.17. A two-tailed t-Test was statistically significant at $p<0.05$.

The findings indicate that the BSM procedure approximated the score on the paper-pencil test for those Group A examinees who took the computer adaptive test procedure first. This group saw only five to nine of the total forty-four items which appeared on the paper-pencil test.

The significantly higher scores on the BSM procedure for Group B examinees, who took the forty-four item paper-pencil test first, were attributable to practice effects occurring between the two test procedures. The time interval for the two procedures for both groups was approximately five weeks.

Supplemental analyses indicate no significant differences between the two groups on their respective test procedures. Table 4.1.2 displays these findings.

Table 4.1.1

Correlated t-Test for Each Group on the Two Different Procedures

		Ability Score			
Group	Procedure	Mean	S.D.	t Value	Significance
A	BSM	95.98	5.49	-0.44	n.s.
	SNAP5	96.16	6.68		
	d.f. = 154				
B	BSM	96.48	5.08	2.22	*
	SNAP5	95.75	5.17		
	d.f. = 156				

* $p<0.05$, n.s. = not significant

Table 4.1.2

t-Test Comparing Groups A and B
on Each Procedure

		Ability Score			
Procedure	Group	Mean	S.D.	t Value	Significance
BSM	A	95.98	5.49	-0.53	n.s.
	B	96.48	5.07		
d.f. = 311					
SNAP5	A	96.16	6.68	0.28	n.s.
	B	95.75	5.17		
d.f. = 311					

n.s. = not significant

Hypothesis H_o:1.2:

There will be no significant relationship between the BSM and SNAP5 ability estimates.

Pearson product-moment correlation coefficients were calculated to determine if a relationship existed for either group between the BSM and SNAP5 ability estimates.

The null hypothesis was rejected for both groups. The correlations were significant at the 0.01 level. Table 4.2.1 displays the findings.

The correlations between the BSM and SNAP5 procedures results for both groups A and B were significant at $p<0.01$. An $r = 0.68$ was reported for Group A, while an $r = 0.69$ was reported for Group B.

Of importance, as seen in Table 4.2.2, was the maximum, minimum, and range of scores of the two groups on both test procedures. The initial-test procedure results for both groups were similar, although the BSM procedure estimates for Group A were consistently lower. These lower estimations were not statistically significant as demonstrated by other analyses found elsewhere in the findings. The range of scores and their proximities at the extremes suggest some potential relationship; no concommitant variable or variables were identified on which to match the groups' examinees. Thus, no test of significance was attempted.

Table 4.2.1

Pearson Product Moment
Correlations Between BSM and
SNAP5 Testing Procedures

Group	N	r	r^2	Significance
A	155	.68	.46	**
B	157	.69	.47	**

** $p<0.01$

Table 4.2.2

Maximum, Minimum, and Ranges of
BSM and SNAP5 Scores

Group	Procedure	Max.	Min.	Range
A	BSM	109.6	79.4	30.2
(N=155)	SNAP5	112.8	69.5	43.3
B	SNAP5	112.8	84.4	28.4
(N=157)	BSM	115.1	81.4	33.7

Hypothesis H_o:2.1:

The use of fifteen to twenty percent of the SNAP5 items will not give a reliable estimate of the examinee's ability.

In order to test this hypothesis, linear regression was employed using the BSM as a predictor of the SNAP5 criterion variable.

The null hypothesis was rejected. The findings indicate a reliable measure of estimated ability for examinees in Group A. The same reliability was indicated for Group B.

The prediction equations for Group A and Group B are listed below, respectively (see Equation 1 and Equation 2). Y' was the SNAP5 criterion variable and X was the BSM predictor variable.

$$Y' = 17.1 + .83X \qquad (1)$$

$$Y' = 28.8 + .69X \qquad (2)$$

Pearson product moment correlation coefficients were calculated for each group. A value of $r = 0.68$ was found for both groups, which was statistically significant at $p<0.01$.

Hypothesis H_o:3.1:

There is no significant difference in mathematical attitude towards testing after the examinee has been exposed to computerized adaptive testing.

In order to study specific changes in mathematics attitude due to exposure to computerized adaptive testing, the twenty mathematics attitude items of the MAS where divided into the ten positive and ten negative response-type

items on the survey. This was done to determine if any significant difference existed due to computerized adaptive testing on either postive-response or negative-response attitude types. A correlated t-test was employed to compare these various mathematics attitudes before and after exposure to computerized adaptive testing within each group.

The hypothesis was not rejected at the 0.05 level of significance for postive nathematics attitudes for either group, nor was it rejected for negative mathematics attitude for Group B at the same significance level. The hypothesis was rejected at the 0.01 level of significance for Group A negative mathematics attitude.A correlated t value of -0.06 was determined for positive mathematics attitude for Group A. A value of 0.41 was calculated using a correlated t-test for positive mathematics attitude for Group B. Neither value was statisically significant at $p<0.05$

A correlated t value of 2.83 was calculated for negative mathematics attitude for Group A. This value was statistically significant at $p<0.01$. The similarly calculated t value of -0.58 for negative mathematics attitude was not significant at $p<0.05$.

The t value of 2.83 demonstrated a significant decline in negative mathematics attitude by examinees in Group A. It was implicit by this that exposure to computerized adaptive testing had a significant effect on negative mathematics attitude for those exposed to the BSM procedure first. The same is not true for Group B examinees who were exposed to the SNAP5 testing procedure first.

Table 4.3.1 and Table 4.3.2 display the findings for this hypothesis.

Table 4.3.1

Correlated t-Test on Positive Mathematics Attitudes Before and After Computer Adaptive Testing

Group	Attitude Measure	Attitude Score Mean	S.D.	t-Value	Significance
A	Before	37.9	8.99	-0.06	n.s.
	After	38.2	9.60		
	d.f. = 154				
B	Before	38.1	9.83	0.41	n.s.
	After	37.9	9.68		
	d.f. = 156				

n.s. = not significant

Table 4.3.2

Correlated t-Test on Negative Mathematics Attitudes Before and After Computerized Adaptive Testing

Group	Attitude Measure	Attitude Score Mean	S.D.	t-Value	Significance
A	Before	20.8	8.04	2.83	**
	After	19.6	8.48		
	d.f. = 154				
B	Before	19.3	7.64	-0.58	n.s.
	After	19.6	8.51		
	d.f. = 156				

** $p<0.01$, n.s. = not significant

Hypothesis H_o:4.1:

The total item response time is not significantly related to the examinee's BSM ability estimate.

A Pearson product moment correlation coefficient was calculated for each group to determine if any significant relationship could be found between total item response time and the examinee's BSM ability estimate. For Group A, a value of $r = 0.11$ with $r^2 = 0.013$ was not significant at $p<0.05$. Likewise, for Group B, $r = 0.03$ with $r^2 = 0.001$ was not significant at $p<0.05$.

The null hypothesis was not rejected at the 0.05 level of significance. Table 4.5 displays the findings.

These findings suggest that individual response time is probably a personal variable independent of the individual's latent ability on the test's dimension. Since the BSM ability estimate is identical to the item difficulty of the highest correctly answered item, it was inferred that the item difficulty and response time were not significantly related. Table 4.4 displays the findings.

Table 4.4

Pearson Product Moment Correlations for Time versus BSM Ability Estimates for Groups A and B

Group	r	r^2	Significance
A (N = 155)	0.11	0.013	n.s.
B (N = 157)	0.03	0.001	n.s.

n.s. = not significant

Hypothesis H_o:4.2:

The total item response time is not significantly related to the examinee's SNAP5 ability estimate.

A Pearson product moment correlation coefficient was calculated for each group to determine if any significant relationship was found between total item response time and the examinee's SNAP5 abiliy estimate. For Group A, a value of $r = -0.09$ with $r^2 = 0.01$ was not significant at $p<0.05$. Similarly, for Group B, a value of $r = -0.13$ with $r^2 = 0.02$ was not significant at $p<0.05$.

The null hypothesis was not rejected at the 0.05 level of significance. Table 4.6 displays the results for both groups.

These findings reflect those of the last hypothetical findings. Table 4.5 displays the findings.

Table 4.5

Pearson Product Moment Correlations for Time versus SNAP5 Ability Estimates for Groups A and B

Group	r	r^2	Significance
A (N = 155)	-0.09	0.009	n.s.
B (N = 157)	-0.13	0.016	n.s.

n.s. = not significant

Hypothesis H_0:4.3:

The total item response time is not significantly related to the examinee's previous mathematics grades.

A Pearson product moment correlation coefficient was calculated for both groups to determine if any significant relationship was found between total item response time and the examinee's previous mathematics grades. For Group A, a value of $r = -0.09$ with $r^2 = 0.008$, was not significant at $p<0.05$. Similarly, for Group B, a value of $r = -0.06$ with $r^2 = 0.004$ was not significant at $p<0.05$.

The hypothesis was not rejected at the 0.05 level of significance. Table 4.6 displays the results for both groups.

Table 4.6

Pearson Product Moment Correlations for Time versus Previous Mathematics Grades

Group	r	r^2	Signifcance
A (N = 155)	-0.09	0.008	n.s.
B (N = 157)	-0.06	0.004	n.s.

n.s. = not significant

Hypothesis H_0:4.4:

> There is no significant relationship between an examinee's SNAP5 ability estimate and the BSM ability estimate, the previous mathematics grades, and the initial-item response time.

In order to test this hypothesis, multiple regression was employed with SNAP5 as the criterion variable and BSM, previous mathematics grades anad initial-item response time at predictor variables. Multiple regression was employed on each group independently.

The null hypothesis was rejected for both groups. For each group, the F-statistic was significant at $p<0.01$. Multiple R values of 0.73 and 0.74 for Groups A and B respectively were also significant at $p<0.01$. Table 4.7 displays the findings.

An examination of the findings indicated a non-significant contribution by the examinee's initial-item response time to predicting the SNAP5 ability estimate. The contribution made by the mathematics grades was significant, but the BSM ability estimate made the major contribution. Both of these contributions were significant at the 0.01 level.

The <u>beta</u> coefficient of 0.57 for the BSM ability estimate demonstrated the major contribution this estimate plays in SNAP5 ability prediction.

Table 4.7

Multiple Regression Analysis
for Groups A and B

Group A (N = 155)

(1) Standardized z-score multiple regression equation:

$$Z'_y = .27z_1 + .57z_2 - .10z_3$$

Z'_y = the predicted SNAP ability estimate

z_1 = examinee mathematics grades

z_2 = BSM ability estimate

z_3 = initial-item response time

(2) ANOVA comparing SNAP5 with BSM, mathematics grades, and item response time

Source	d.f.	S.S.	M.S.	F	Sig.
Regression	3	3649.63	1216.54	56.88	*
Residual	151	3229.62	21.39		

* $p<0.01$

(3) Multiple R = 0.73

R Squared = 0.53

(4) Summary:

Variable	Multiple R	R^2
Mathematics Grades	0.50	0.25
BSM	0.72	0.52
Time	0.73	0.53

Table 4.7(continued)

Group B (N = 157)

(1) $Z'_y = .26z_1 + .59z_2 - .13z_3$

(2) ANOVA comparing SNAP5 with BSM, mathematics grades, and item response time

Source	d.f.	S.S.	M.S.	F	Sig.
Regression	3	2299.57	766.52	62.66	*
Residual	153	1871.56	12.23		

* $p<0.01$

(3) Multiple R = 0.74

R Squared = 0.55

(4) Summary:

Variable	Multiple R	R^2
Mathematics Grades	0.43	0.18
BSM	0.73	0.53
Time	0.74	0.55

Hypothesis H_o:4.5:

There is no significant relationship between the SNAP5 ability estimate, and the BSM ability estimate, and the examinee's previous mathematics grades

In order to study this hypothesis, multiple regression was employed independently on both groups with SNAP5 as the criterion variable, and BSM and mathematics grades as predictor variables.

The hypothesis was rejected for both groups. Mulitple R values of 0.72 and 0.73 for Groups A and B respectively were significant at $p<0.01$. Also, the F ratios for both groups were significant at $p<0.01$. Table 4.8 displays the findings.

An examination of the findings shows the significant contribution by both independent variables to prediction of the SNAP5 ability estimate. The major contribution is made by the BSM ability estimate to prediction of the SNAP5 ability estimate.

Table 4.8

Multiple Regression Analysis
on Both Groups

Group A (N = 155)

(1) Standardized z-score multiple regression equation

$Z'_y = .27z_1 + .57z_2$

where,

Z'_y = predicted SNAP5 ability estimate

z_1 = previous math grades

z_2 = BSM ability estimate

(2) ANOVA of SNAP5 with BSM and Mathematics Grades

Source	d.f.	S.S.	M.S.	F	Sig.
Regression	2	3580.02	1790.01	82.47	*
Residual	152	3299.23	21.71		

* $p<0.01$

(3) Multiple R = 0.72

R Squared = 0.52

(4) Summary:

Variable	Multiple R	R Squared
Mathematics Grades	0.50	0.25
BSM	0.72	0.52

Table 4.8(continued)

Group B (N = 157)

(1) $Z'_y = .27z_1 + .61z_2$

(2) ANOVA of SNAP5 with BSM and Mathematics Grades

Source	d.f.	S.S.	M.S.	F	Sig.
Regression	2	2228.91	1114.46	88.37	*
Residual	154	1942.21	12.61		

* $p<0.01$

(3) Multiple R = 0.73

R Squared = 0.53

(4) Summary:

Variable	Multiple R	R Squared
Mathematics Grades	0.53	0.18
BSM	0.73	0.53

Hypothesis H_o:4.6:

There is no significant relationship between the SNAP5 ability estimate, and the highest correctly answered BSM item difficulty value

In order to study this hypothesis, multiple regression was employed with SNAP5 as the criterion variable and BSM as the predictor variable.

The hypothesis was rejected for both groups. The multiple R values for Groups A and B were both 0.68 and were significant at $p<0.01$. The F ratio for both groups was also significant at $p<0.01$. Table 4.9 displays the findings.

These findings are consistent with the findings of the two previous hypotheses. This consistency lends credence to the predictive value of the BSM procedure.

Table 4.9

Multiple Regression Analysis
on Both Groups

Group A (N = 155)

(1) Standardized z-score multiple regression equation

$Z'_y = .68z$

where,

Z'_y = predicted SNAP5 ability estimate

z = BSM ability estimate

(2) ANOVA on SNAP5 with BSM

Source	d.f.	S.S.	M.S.	F	Sig.
Regression	1	3155.43	3155.43	129.65	*
Residual	153	3723.82	24.34		

* $p<0.01$

(3) Multiple R = 0.68

R Squared = 0.46

Table 4.9(continued)

Group B (N = 157)

(1) $Z'_y = 0.68z$

(2) ANOVA on SNAP5 with BSM

Source	d.f.	S.S.	M.S.	F	Sig.
Regression	1	1940.72	1940.72	134.87	*
Residual	155	2230.40	14.39		

* $p<0.01$

(3) Multiple R = 0.68

R Squared = 0.46

Hypothesis H_o:4.7:

> There is no significant relationship between the BSM ability estimate and the SNAP5 ability estimate, and the examinee's previous mathematics grades.

In order to study this hypothesis, multiple regression was employed with BSM as the criterion variable and SNAP5 and mathematics grades as predictor variables.

The hypothesis was rejected for both groups. The multiple R values for both groups were 0.68, which were significant at $p<0.01$. The F ratios for both groups were also significant at $p<0.01$. Table 4.10 displays the findings.

An examination of the findings shows a non-significant contribution to prediction of the BSM ability estimate made by the previous mathematics grades. The contribution made by the SNAP5 ability estimate was significant.

These findings coincide with the previous hypothesized findings. Prediction of one testing procedure by another provides for a measure of linearity of the prediction equations. Further, the major contributions in predicting either score are by the other testing procedure score.

Table 4.10

Multiple Regression Analysis
of Both Groups

Group A (N = 155)

(1) Standardized z-score multiple regression equation

$Z'_y = .08z_1 + .64z_2$

where,

Z'_y = predicted BSM ability estimate

z_1 = previous math grades

z_2 = SNAP5 ability estimate

(2) ANOVA of BSM with SNAP5 and Mathematics Grades

Source	d.f.	S.S.	M.S.	F	Sig.
Regression	2	2156.41	1078.21	65.78	*
Residual	152	2491.46	16.39		

* $p<0.01$

(3) Multiple R = 0.68

R Squared = 0.46

(4) Summary:

Variables	Multiple R	R Squared
Mathematics grades	0.40	0.16
SNAP5	0.68	0.46

Table 4.10(continued)

Group B (N = 157)

(1) $Z'_y = -.05z_1 + 70z_2$

(2) ANOVA of BSM with SNAP5 and Mathematics Grades

Source	d.f.	S.S.	M.S.	F	Sig.
Regression	2	1877.89	938.94	76.47	*
Residual	154	2143.23	13.92		

* $p<0.01$

(3) Multiple R = 0.68

R Squared = 0.46

(4) Summary:

Variable	Multiple R	R Squared
Mathematics grades	0.25	0.06
SNAP5	0.68	0.46

Hypothesis H_o:5.1:

> There will be no significant attitude differences between the two groups before, between or after the two ability estimation procedures.

In order to test this hypothesis, student attitudes at the three data points, before, during and after, were compared using correlated t-test. Group A and Group B were analyzed separately. The results are recorded in Table 4.11.

The hypothesis was not rejected at the 0.05 level of significance. Examination of these results indicates that no significant differences were found between Before and During, During and After, and Before and After comparisons for either group. All of the t values were between -0.6 and 1.29. The means were similarly homogeneous.

Table 4.11

Correlated t-Tests on
Total Mathematics Attitude
Measures Before, During and After
Both Testing Procedures

Attitude Measure	Mean	S.D.	t-Value	Significance
Group A:				
Before	37.99	8.99	-0.60	n.s.
During	38.24	9.60		
During	38.24	9.60	1.17	n.s.
After	37.72	10.12		
Before	37.99	8.99	0.58	n.s.
After	37.72	10.12		
d.f. = 154				
Group B:				
Before	38.56	9.28	0.98	n.s.
During	38.12	9.83		
During	38.12	9.83	1.29	n.s.
After	37.92	9.68		
Before	38.56	9.28	1.29	n.s.
After	37.92	9.68		
d.f. = 156				

n.s. = not significant

Unhypothesized Results

There were two post hoc analyses which related significantly to the postulated results of this study. The first result dealt with how well the highest correct item (HCI) of the BSM test procedure approximated the Rasch ability estimate as determined by the SNAP5 paper-pencil test procedure. The relatve efficiency function, $I = 400P(1-P)$, where P is the probability of a correct response, was used to determine this HCI fit. The second result was concerned with using the two test procedures independently to ascertain a pass or fail status on the test itself.

The probability of a correct response, P, is a function of the difference between the examinee ability estimate and the difficulty of the item encountered by the examinee. In determining the relative efficiency, I, the ability estimate was determined by using the SNAP5 raw score and the encountered item difficulty by the HCI of the BSM procedure. The SNAP5 ability was considered the standard ability estimate against which the BSM estimate was compared.

Table 4.12 displays results related to the first post hoc analysis. The results are separated by group because of the different test conditions for Groups A and B. The results presented in the table indicate that the BSM HCI did well in approximating the SNAP5 paper-pencil test score. For instance, an information value of 90% or more indicates an absolute difference between estimated ability and item difficulty of less than 0.7 logits. The table displays that

over 87% of the examinees from both groups were in this category. This information cutoff of 90% also indicates that it is very difficult to detect item misfit from the examiee's latent ability measure on this test's dimension.

The subject of the second post hoc analysis was a consideration for the BSM procedure to determine pass/fail on the test in a similar manner to the SNAP5 procedure. The cut score used for a pass/fail determination was the grand mean of the two procedures minus one standard deviation (i.e., mean = 96.1, std. dev. = 55.6). The frequencies for pass/fail across both procedures are contained in the 2 x 2 contingency table of Table 4.13.

If no significant differnce among the cells of the table was found, then it must be concluded that the BSM procedure did not distinguish a pass/fail condition analogous to the standard set with the SNAP5 procedure. That is, the BSM did not perform in a similar capacity.

A 2 x 2 Chi-square test was performed on the dichotomized data. The Chi-square = 35.64 which was significant at $p<0.01$. This means that the two test procedures results in equivalent pass/fail decisions for all students. Therefore, the BSM procedure did equally well as did SNAP5 in distinguishing pass/fail on the test.

Table 4.12

Frequency Distribution of
Relative Information Percentages
for Groups A and B

Percentage Range	Group A	Group B
0 - 9	0	0
10 - 19	1	0
20 - 29	0	0
30 - 39	1	0
40 - 49	0	0
50 - 59	1	1
60 - 69	1	1
70 - 79	6	8
80 - 89	9	8
90 - 99	136	139
	N = 155	N = 157

Table 4.13

Chi-square 2 x 2 Frequency Table and Statistics for Pass/Fail Determination

		SNAP5		
		Fail	Pass	
BSM	Fail	20	24	44
	Pass	26	242	268
		46	266	312

Chi-square = 35.64 with d.f. = 1

Phi coeff. = 0.35

CHAPTER V

CONCLUSIONS, LIMITATIONS AND RECOMMENDATIONS

Chapter V includes a discussion of the conclusions, limitations and recommendations based on the findings of the study. In summary, two independent samples of fifth grade students were administered two different testing procedures, a paper-pencil test (SNAP5) and the BSM computer adaptive test. Both procedures utilized the same item bank. Group A examinees were given the BSM procedure first, while examinees in Group B were given the SNAP5 test first.

For each group, correlations and scattergrams of individual classes, correlated t-Tests and multiple regression analyses were performed on the two testing procedures. In addition, correlated t-Tests were performed on the attitude surveys given before, between and after the two testing procedures. Findings indicated no significant differences in ability estimates when comparing the two procedures. When looking at prediction of an ability estimate, either by the SNAP5 or BSM procedures, the strongest contribution was made by either testing procedure.

Conclusions

The BSM procedure showed a strong correlation with and good prediction of the SNAP5 procedure of ability estimate. Multiple regression analyses demonstrated repeatedly that the strongest contribution to prediction was by the BSM ability estimate. The prediction to the SNAP5 estimate by

the BSM procedure lost some of its accuracy by the varying testing milieu which existed as an uncontrollable contaminant in the various classes. This milieu was largely, if not solely, determined by the classroom teacher. This was demonstrated by the consistency of ability estimate made by the BSM procedure, and the inconsistency of ability estimate made by the SNAP5 test. In those classes which showed a low correlation between the two procedure estimates, the testing milieu was markedly poor. The correlations between the two procedures for all twenty classes ranged from 0.23 to 0.87. The individual group correlations and their respective scattergrams appear in Appendix B. From these statistics it appears that the teachers influence in the testing milieu had a significant influence on the performance of an examinee on a paper-pencil test. This assertion was evidenced by an informal survey of these teachers by the author. It was determined that for those teachers who established a good testing environment also demonstrated a high correlation between the results of the two procedures.

It was found that a proper, teacher-developed testing environment was not necessary when the BSM procedure was employed. The computerized testing procedure caused the necessary and sufficient testing environment for the examinee. It controlled the presentation rate of the items, focused the attention of the examinee to the task at hand, demonstrated its "patience" by allowing ample time for response and focused the examinee's attention to a single test item.

In the attitude survey information, there was no significant difference before, between and after either testing procedure. This relates well to finding that there was little or no correlation between positive mathematics attitude and either of the ability estimation procedures.

The total amount of response time on the computerized procedure demonstrated no correlation between either of the ability estimating procedures. It was initially hypothesized that a strong negative correlation would be demonstrated, but this was not substantiated by the data. It would appear that the amount of time an examinee takes to respond to an item may be dependent on some unhypothesized variable or variables not accounted for in this study.

The finding that no significant relationship exists between total item response time and BSM ability estimate further underscores the importance that time not be used as a constraint when certain objective test measurements are made. The examinees perform nearly as well when provided with fewer test items and virtually unlimited response time.

Whether or not time factors of individual response set are significantly related with conventional paper-pencil tests in general was not determined. The length of time spent by each examinee on the SNAP5 procedure was not one of the data points collected during this study.

The results of this study showed a statistically significant relationship between the BSM and SNAP5 ability estimates. Although the reliability coefficient was not exceedingly high, there was support for the assumption that

the BSM procedure produced a reasonably accurate estimate of the SNAP5 ability measure. This reasonably accurate estimate was made using only fifteen to twenty percent of the available item pool anad was considered a valid finding.

Among the 312 examinees who were administered the BSM procedure, 34, or slightly less than eleven percent, responded to additional items beyond the normal five or six items. This occurred because the BSM algorithm determined a Condition I or Condition II item sequencing for these examinees. As described in Chapter III, Condition I was a false negative response on the initial response, while Condition II was a false positive one. The BSM and SNAP5 ability estimates for these particular examinees were compared separately for significant disparity between their scores. This comparison demonstrated virtually the same correlation between the ability estimates of this subgroup as the correlation between the estimates of the entire sample. Also, it was of interest that the percentage of false negatives and positives in the Monte Carlo simulation was virtually identical to that of the actual field study.

The validity of this study was supported in part by the post hoc analysis of the relative efficiency of the HCI to estimate the SNAP5 estimate of ability. The large percentage of high efficiency demonstrates that the BSM algorithm sequenced the items in such a way as to arrive at an item whose difficulty level was analogous to the latent ability of the examinee. One of the salient features of the Rasch model is that person ability measurement is best when item

difficulty and person ability parameters are equal or nearly so. Concomitantly, as the relative efficiency approaches one hundred percent, the ability and difficulty parameters near equality.

Since an examinee's previous mathematics grade had little predictive value to either the SNAP5 or BSM ability estimates, its use as prior information for initializing the sequencing the items was seen as minimal. The fact that the BSM procedure significantly approximated the SNAP5 ability estimate without use of any prior information is of value to the conclusions of this study.

Limitations

The conclusions of this study were limited in several ways. First, the testing was done in the fall of the school year when the examinees had received very little instruction on the material included on the test. This minimal learning led to frustration in varying degrees on the part of the examinees. The teachers, acting as test proctors, reacted in a variety of ways to these frustrations. Their reactions resulted, in some instances, in changing the testing milieu.

Second, the item bank was limited both by the number of items and by the upper and lower bounds of the item difficulties. For this sequencing and selection algorithm, at least one hundred items should be included, and the upper and lower item difficulty bounds should be +3.0 and -3.0 Rasch units respectively.

Finally, the sequencing and selection algorithm could be rewritten in an alternate form to include item

discrimination indices. These two computer adaptive testing procedures, the original BSM and this altered version, could be compared on this added dimension.

Recommendations

The field of computerized adaptive testing is relatively new, especially when applied with microcomputers. The selection and secquencing algorithms that exist are many and varied as well as being based on a one- or three-parameter latent trait model. The item bank size and dimensions are also of theoretical concern.

Considerable research is needed in the area of microcomputer-based computerized adaptive testing. Research efforts should be aimed at solving the problems of how many parameters should be included for use with micro-computerized adaptive testing; the optimal size of the adaptive testing item bank; and the sufficiency of the computerized adaptive testing selection and sequencing algorithm.

Additional research should address the above research issues in various subject areas of the schools. There are different concerns for language arts, science, more advanced mathematits, social studies, and other curriculum areas regarding micro-computerized adaptive testing.

It will be through additional research in the aforementioned areas and others that the full realization of micro-computerized adaptive testing will occur. More generalizable procedures need to be formulated and tested through various research efforts. This study has provided a

necessary step toward the realization of more objective and valid testing procedures for which micro-computerized adaptive testing offers considerable hope.

Problems for Further Study

The following guidelines are given for use in the future and much needed research in the area of computerized adaptive testing. These guidelines should be used to control largely for the effects of local test proctor influence on examinees.

1. Two independent groups should be determined and be given only the computerized adaptive test procedure. One or more concommitant variables should be used to match subjects before looking for significant differences in ability estimation. The use of the same test administered by a paper-pencil ability estimation procedure leaves too much room for error variance as noted above.

2. The two test procedures should be administered to the selected sample in the early spring of the school year to allow the examinees an opportunity to achieve some learning of the relavant material. This could also provide the concommitant variable for matching, as teachers have gathered sufficient grading data by this time.

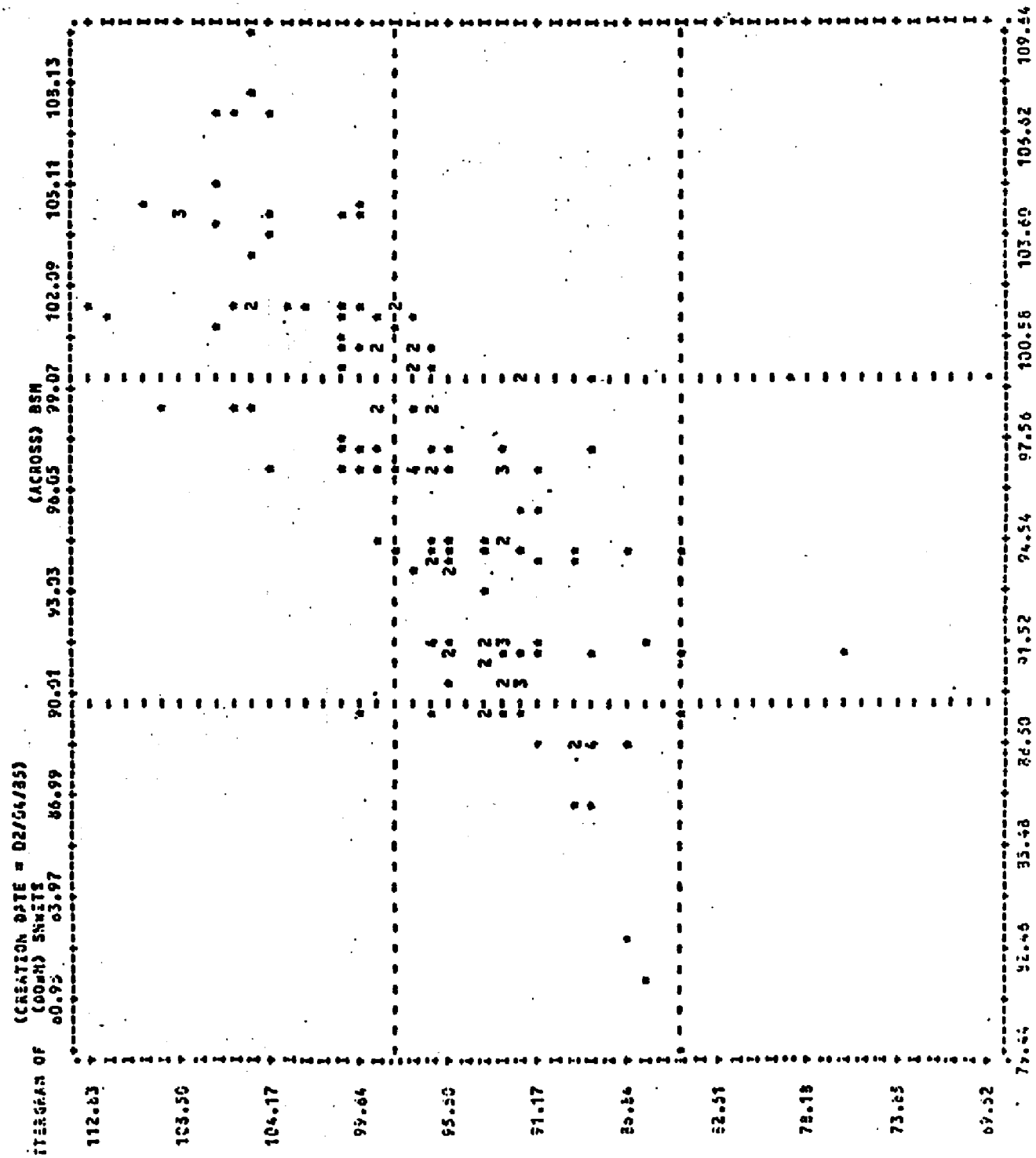
3. Attitude surveys serve a useful function in so much as they allow the examinee to express his views on school and education in general. Additional or replacement attitude items should consist of those dealing with feelings about his taking tests, teacher, school and others.

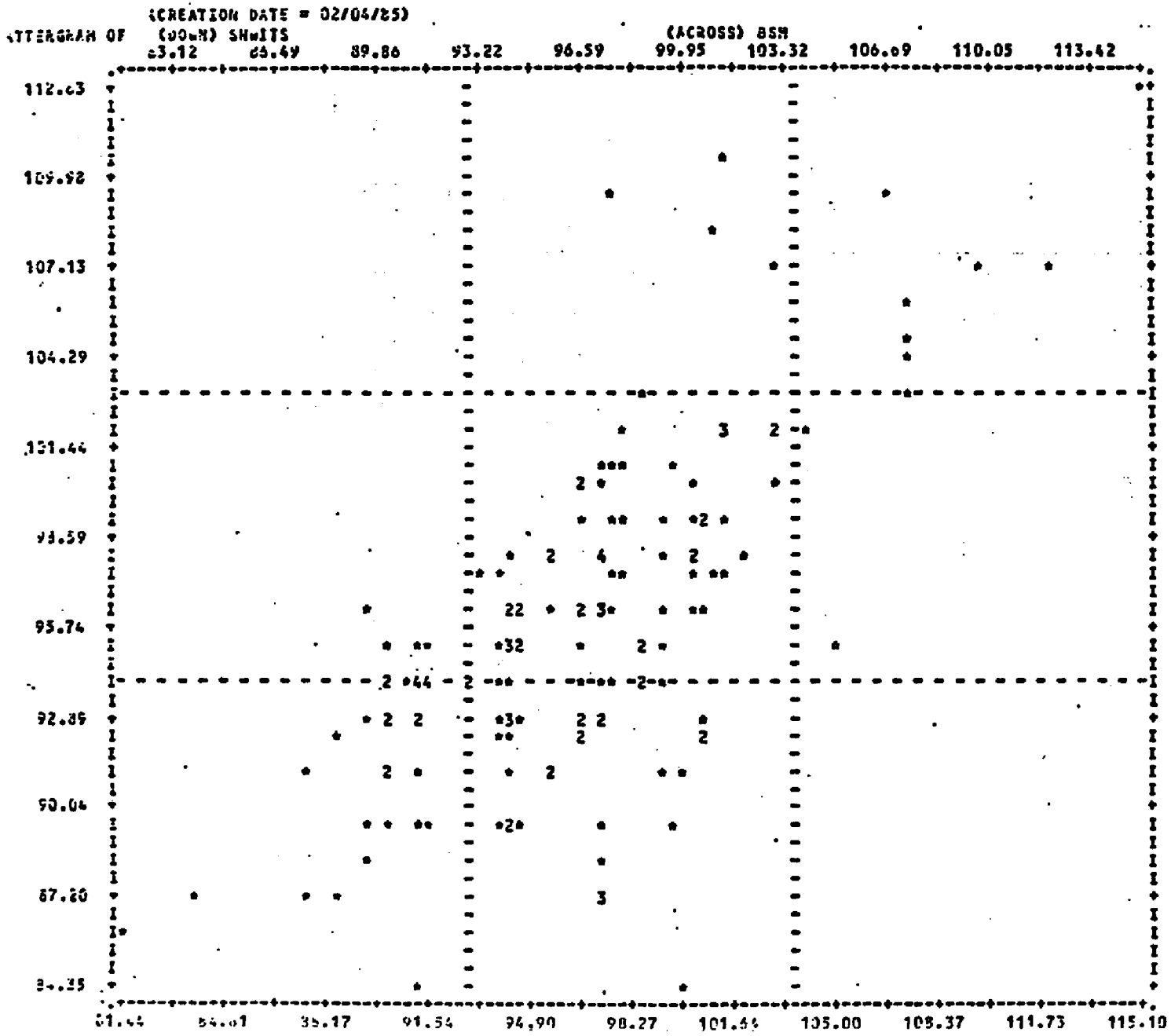
Appendices

Appendix A

Scattergrams and Their Respective
Pearson Product Moment
Correlation Coefficients
for Group A and Group B

Group A
N = 155 r = 0.68



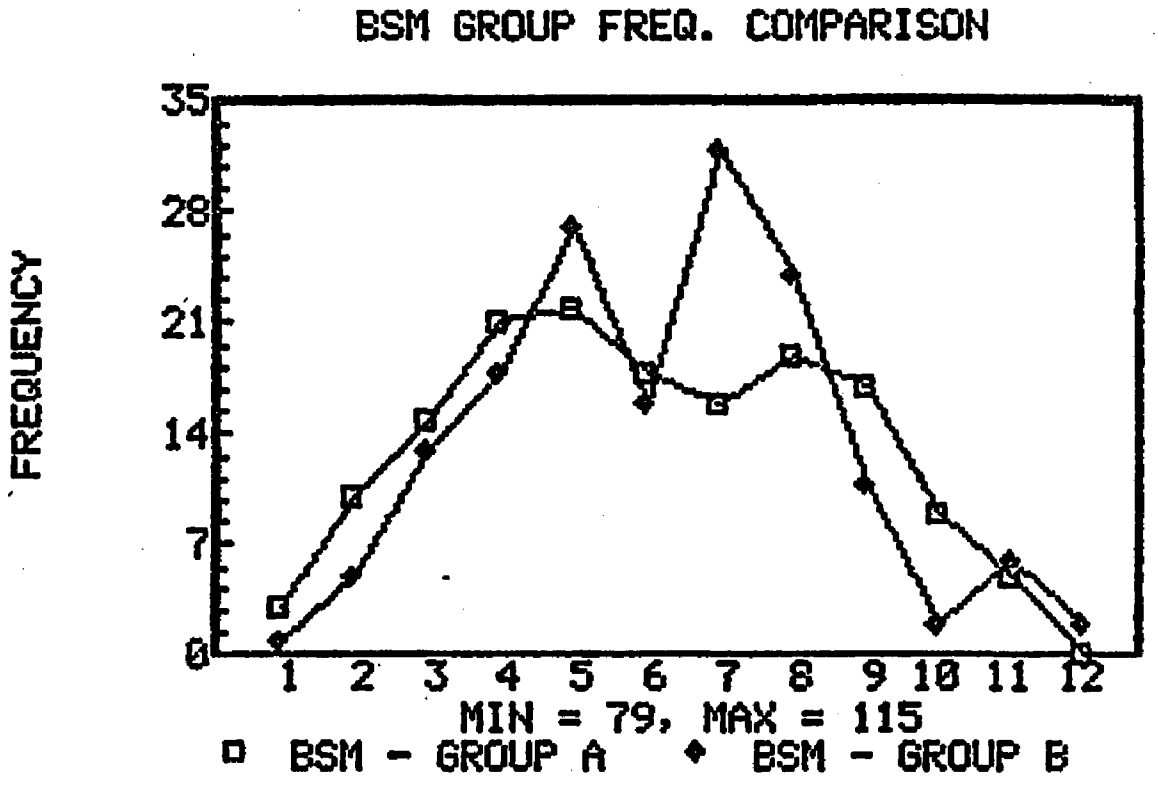


Group B
N = 157
r = 0.68

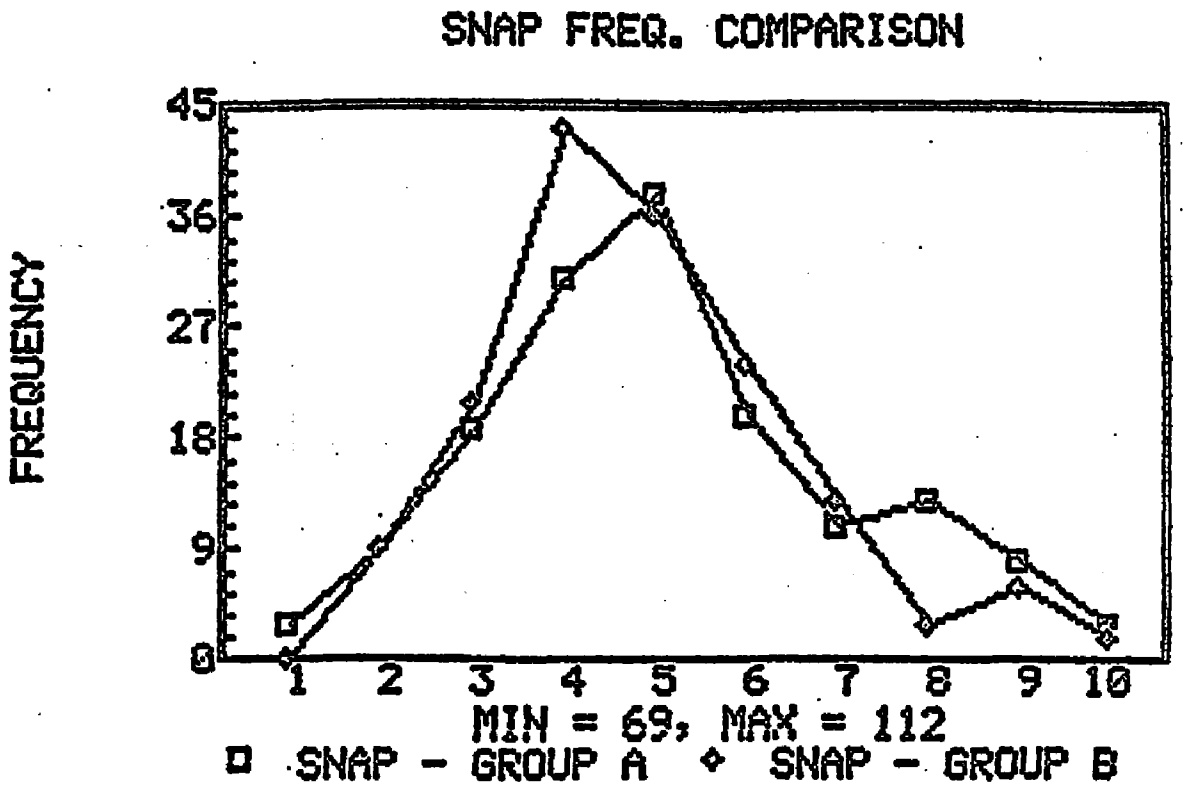
Appendix B

Frequency Distributions of
BSM and SNAP5 WITS Scores
Between and Within the Two
Testing Procedures

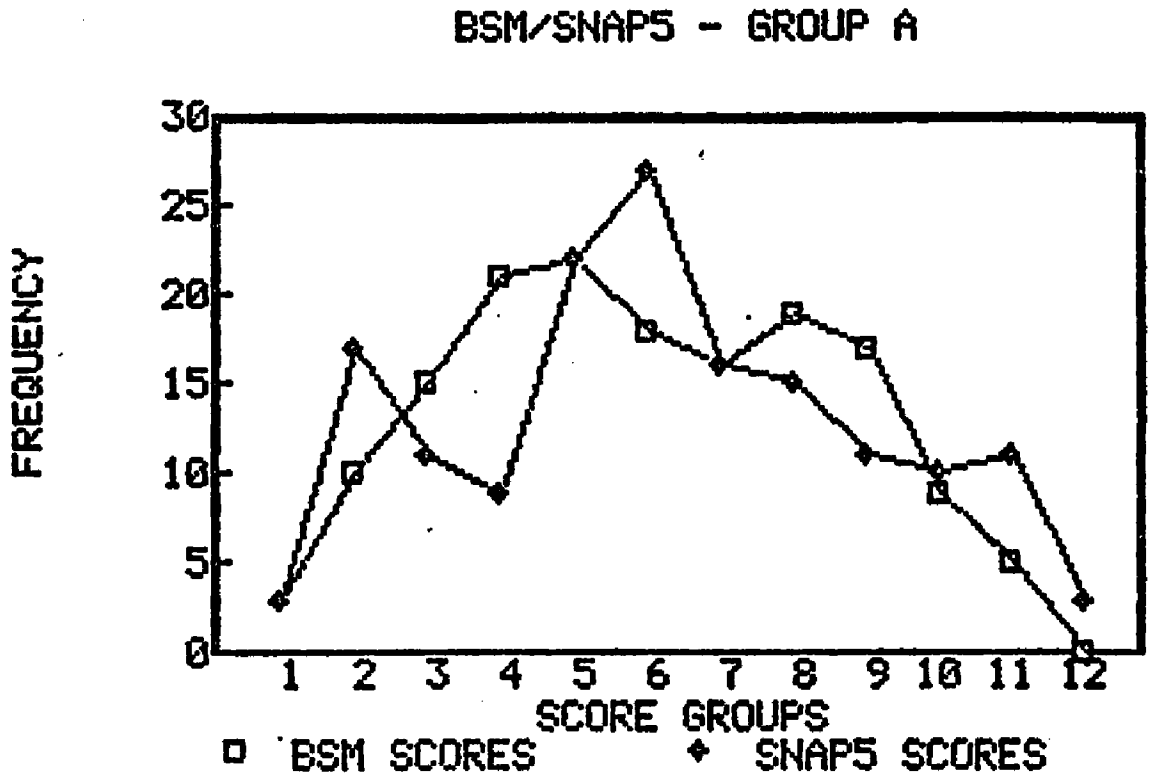
Frequency Distribution of BSM Scores
- A Comparison of Group A and Group B



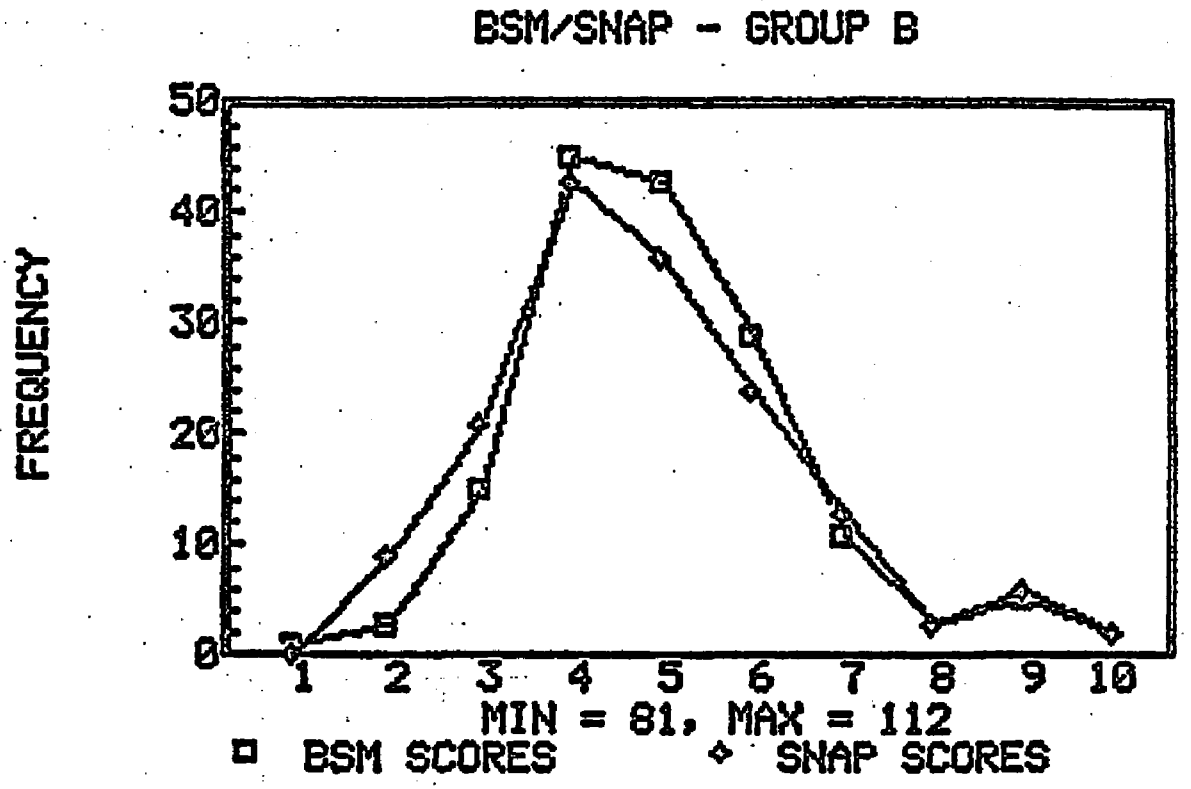
Frequency Distribution of SNAP5 Scores
- A Comparison of Group A and Group B



Frequency Distribution of BSM
and SNAP5 Scores
- A Comparison for Group A



Frequency Distribution of BSM
and SNAP5 Scores
- A Comparison for Group B



BIBLIOGRAPHY

Andersen, E.B. "Asymptotic properties of conditional maximum likelihood estimators". Journal of the Royal Statistical Society B, 1970, 32.

Andersen, E.B. "Asymptotic properties of conditional likelihood ratio sets". Journal of the American Statistical Association, 1971, 66.

Barndorff-Nielsen, O. Information and Exponential Families in Statistical Theory. New York: John Wiley and Sons, 1978.

Benson, J. "A comparison of the one- and three-parameter logistic models on measures of test efficiency." Educational Research Quarterly, Summer, 1982, v.7, n.2

Birnbaum, A. "Some latent trait models and their use in inferring an examinee's ability." In F.M. Lord & M.R. Novick, Statistical Theories of Mental Test Scores. Reading, Mass.: Addison-Wesley, 1968.

Brink, N. "Rasch's logistic model versus the Guttman model." Educational and Psychological Measurement, 1972, 32.

Cook, L. & Hambleton, R. "Latent trait models and their use in the analysis of educational test data." Journal of Educational Measurement, 1977, 14.

Cudek, R., McCormick, D., & Cliff, N. "Implied orders tailored testing[simulation with the Stanford-Binet". Applied Psychological Measurement, Spring, 1980, v.4, n.2.

Edwards, J. "Computer-assisted testing." Journal of Research and Development in Education, 1980, v.14, n.1.

Forster, F. "The optimum size for Rasch field testing." NorthWest Evaluation Association Journal, May, 1977, v.1, n.2.

Forster, F. & Doherty, V. "Using the Rasch approach' to measurement to solve practical school testing problems." Paper presented at the annual meeting of the American Educational Research Association, Toronto, March 1978.

Green, B. "Comments on tailored testing." In W.H. Holtzman (Ed.) Computer-Assisted Instruction, Testing, and Guidance. New York: Harper and Row. 1970.

Green, B., et. al. "Evaluation plan for the computerized adaptive vocational aptitude battery." Dept. of Psychology, Johns Hopkins University, Baltimore, MD. May, 1982.

Gustafsson, J. "An introduction to Rasch's measurement model." National Swedish Board of Education, Stockholm, June, 1980.

Hansen, D. "An investigation of computer-based science testing." In R.C. Atkinson and H.A. Wilson (Ed.). Computer-Assisted Intruction. New York: Academic Press, 1969.

Jelden, D. "Computer generated testing." AEDS Monitor. v.20, n.7,8,9, 1982.

Katz, L. & Dalby, J. "Computer-assisted and traditional psychological assessment of elementary-school-aged children." Contemporary Educational Psychology. 6, October, 1981.

Kingsbury, G. & Weiss, D. "An alternate-forms reliability and concurrent validity comparison of Bayesian adaptive and conventional ability test." (Research Report 80-5). Dept. of Psychology, Univ. of Minnesota, Minneapolis, Dec., 1980.

Loevinger, J. "Person and population as psychometric concepts". Psychological Review, 1965, 72.

Lord, F. "Practical applications of item characteristic curve theory." Journal of Educational Measurement. 14, 1977.

Lord, F. "Some test theory for tailored testing." In W.H. Holtzman (Ed.), Computer-Assisted Instruction, Testing, and Guidance. New York: Harper and Row, 1970.

Lord, F. & Novick, M. Statistical Theories of Mental Test Scores. Reading, Mass.: Addison-Wesley, 1968.

Lord, F. "Tailored testing: An application of stochastic approximation." Journal of the American Statistical Association. v.66, n.336, Dec., 1971.

Lord F. A Theory of Test Scores. Princeton, New Jersey: Educational Testing Service, 1952.

Lord, F. Applications of Item Response Theory to Practical Testing Problems. Hillsdale, NJ: Erlbaum, 1980.

McBride, J. "Adaptive mental testing: the state of the art." Army Research Institute for Behavorial and Social Sciences, Alexandria, VA., Nov., 1979.

McKinley, R. & Reckase, M. "A comparison of a Bayesian and a maximum likelihood tailored-testing procedure." Tailored Testing Research Laboratory. Univ. of Missouri, Columbia, MO. June, 1981.

McKinley, R. & Reckase, M. "Computer applications to ability testing." AEDS Journal. v.13, n.3, Spring, 1980.

McKinley, R. & Reckase, M. "A successful application of latent trait theory to tailored testing." (Research Report 80-1). Educational Psychology Dept., Uni. of Missouri, Columbia, MO. Feb., 1980.

Mead, R. "Using the Rasch model to identify person-based measurement disturbances." In D. Weiss (Ed.). Proceedings of the 1979 Computerized Adaptive Testing Conference. Univ. of Minnesota at Minneapolis, Wayzata, Minn. June, 1979.

Nicewander, W. "A comparison of two methods of interactive testing: final report." Dept. of Psychology, Uni. of Oklahoma, Norman, OK. June, 1980.

Owen, R. "A Bayesian sequential procedure for quantal response in the context of adaptive mental testing." Journal of the American Statistical Association. 70, 1975.

Rasch, G. Probabilistic Models for Some Intelligence and Attainment Tests. Copenhagen: Danmarks Paedogogiske Institut, Nielsen & Lydiche, 1960.

Rasch, G. "An item analysis which takes individual differences into account. British Journal of Mathematical and Statistical Psychology. v.19, n.1, 1966.

Reckase, M. "Tailored testing, measurement problems, and latent trait theory." Paper presented at the annual meeting of the National Council on Measurement in Education. April, 1981.

Robbins, H. & Monro, S. "A stochastic approximation method." The Annals of Mathematical Statistics, 22, Sept., 1951.

Rosso, M. & Reckase, M. "A comparison of a maximum likelihood and a Bayesian ability estimation procedure for tailored testing." Uni. of Missouri, Columbia, MO., April, 1981.

Rushby, N. Computers in the Teaching Process. New York: John Wiley & Sons, 1979.

Sampson, J. "Computer-assisted testing and assessment: current status and implications for the future." Measurement and Evaluation in Guidance. v.15, n.3, Jan., 1983.

Schmidt, F., Urry, V., & Gugel, J. "Computer assisted tailored testing: examinee reactions and evaluations." Educational and Psychological Measurement. 38, 1978.

Spineti, J. & Hambleton, R. "A computer simulation study of tailored testing strategies for objective-based instructional programs." Educational and Psychological Measurement, 37, 1977.

Sympson, J. & Weiss, D. "Predictive validity of conventional and adaptive tests in an Air Force training environment." Dept. of Psychology, Univ. of Minnesota, Minneapolis, MN. Aug., 1982.

Thompson, J. & Weiss, D. "Criterion-related validity of adaptive testing strategies." (Research Report 80-3). Uni. of Minnesota, Minneapolis, MN., June, 1980.

Urry, V. "Tailored testing: a successful application of latent trait theory." Journal of Educational Measurement. 14, Summer, 1977.

Waters, B. "An empirical investigation of the stratified adaptive computerized testing model." Applied Psychological Measurement. v.1, n.1, 1977.

Weiss, D. "Improving measurement quality and efficiency with adaptive testing." Applied Psychological Measurement. v.6, n.5, Fall, 1982.

Wetzel, C. & McBride, J. "Influence of fallible item parameters on test information during adaptive testing." Navy Personnel Research and Development Center, San Diego, CA. April, 1983.

Wright, B. "Solving measurement problems with the Rasch model." Journal of Educational Measurement. 14, Summer, 1977.

Wright, B & Mead, R. "CALFIT: Sample-free item calibration with a Rasch measurement model." Research Memorandum No. 18, Uni. of Chicago. March, 1975.

Wright, B & Stone, M. Best Test Design. Chicago.: Mesa Press. 1979.

Wright, B., Mead, R., & Bell, S. "BICAL: Calibrating items with the Rasch model." (Statistical Laboratory Research Memorandum No. 23B). Chicago: University of Chicago, Dept. of Education, June 1979.

ABSTRACT

AN APPLICATION OF THE RASCH MODEL TO COMPUTERIZED ADAPTIVE TESTING: THE BINARY SEARCH METHOD

by

DENNIS RUSSELL WISNIEWSKI

December, 1985

Adviser: Donald Marcotte, Ph.D.

Major : EVALUATION AND RESEARCH (EDUCATION)

Degree : DOCTOR OF PHILOSOPHY

The purpose of the research was to determine if the Binary Search Method (BSM) of computerized adaptive testing could accurately estimate an examinee's ability; to study the effects on attitudes of examinees towards computerized adaptive testing and conventional paper-pencil testing procedures; and to study the relationship between item response time and item difficulty. A union of increasingly sophisticated microcomputer technology and adaptive testing procedures based on latent trait theory have provided the essentials for an exminiation of individualized testing.

The research involved the viability of the Binary Search Method as a tool for providing an accurate estimate of an examinee's ability using a minimum of test items. The BSM procedure showed a high correlation and was a valid predictor of the paper-pencil results of ability estimate. Multiple regression analyses demonstrated that the strongest contribution to prediction was the BSM ability estimate.

The prediction of the paper-pencil estimate by the BSM procedure lost some of its accuracy due to varying paper-pencil testing milieu which existed as an uncontrollable contaminant in the various classes. It was significant that such a prior testing environment was not necessary when the computer procedure was employed. It appeared that the computerized testing procedure created a necessary and sufficient testing environment for the examinee. It controlled the presentation rate of the items, focused the attention of the examinee to the task at hand, demonstrated the need for patience in test item presentation, and focused the examinee's attention to a single problem.

The total amount of response time using the computerized procedure was not correlated to either of the ability estimating procedures. It was initially hypothesized that a highly negative correlation would be found, but this was not substantiated by the data.

Future research efforts should be aimed at solving the problems of which parameter-size model is best suited with micro-computerized adaptive testing; what is the optimal size of the adaptive testing item bank; and how efficient is the computerized adaptive testing selection and sequencing algorithm.

AUTOBIOGRAPHICAL STATEMENT

DENNIS RUSSELL WISNIEWSKI

Born: August 20, 1944
Detroit, Michigan

Education:

Pershing High School, Detroit, Michigan

Mira Costa Community College, Oceanside, California

Wayne State University, Detroit, Michigan
Bachelor of Science, 1970
Master of Arts in Mathematics Education, 1975
Doctor of Philosophy in Evaluation and Research (Education), 1985

Military Service:

U.S. Marine Corps, Fleet Marine Force, Pacific, 1963-1966

Special Honors:

Wayne State University Graduate/Professional Scholarship, (1978-1979)

Professional Presentations:

"Tailored Testing on Micro-computers:
A local application of the Rasch model"
Presented at the annual meeting of the Michigan Educational Research Association, Ann Arbor, MI, January, 1985.

Professional Memberships:

American Eductional Research Association
American Statistical Association
National Council on Measurement in Education
Michigan Educational Research Association

Employment:

Teacher, Friends School in Detroit, Detroit, Michigan, 1970-1971
Teacher, Utica Community Schools, Utica, Michigan, 1971-1984
Supervisor of Evaluation, School District of the City of Pontiac, Pontiac, Michigan, 1984 to present