1-1-1973

Computer simulation study of tailored testing strategies for objective-based instructional programs.

John P. Spinetti
University of Massachusetts Amherst

Follow this and additional works at: https://scholarworks.umass.edu/dissertations_1

Recommended Citation

Spinetti, John P., "Computer simulation study of tailored testing strategies for objective-based instructional programs." (1973). *Doctoral Dissertations 1896 - February 2014*. 2908.
<https://doi.org/10.7275/q16f-bp94> https://scholarworks.umass.edu/dissertations_1/2908

This Open Access Dissertation is brought to you for free and open access by ScholarWorks@UMass Amherst. It has been accepted for inclusion in Doctoral Dissertations 1896 - February 2014 by an authorized administrator of ScholarWorks@UMass Amherst. For more information, please contact scholarworks@library.umass.edu.

UMASS/AMHERST

312066013583534

COMPUTER SIMULATION STUDY OF TAILORED TESTING
STRATEGIES FOR OBJECTIVE-BASED INSTRUCTIONAL PROGRAMS

A Dissertation Presented

by

JOHN P. SPINETI

Submitted to the Graduate School of the
University of Massachusetts in
partial fulfillment of the requirements for the degree of

DOCTOR OF EDUCATION

November 1973

Educational Research--Science Education

(c) John P. Spinetti 1973

All Rights Reserved

A COMPUTER SIMULATION STUDY OF TAILORED TESTING
STRATEGIES FOR OBJECTIVE-BASED INSTRUCTIONAL PROGRAMS

A Dissertation

by

JOHN PATRICK SPINETI

Approved as to style and content by:

Dr. Ronald K. Hambleton (Chairman of Committee)

Dr. Leverne J. Thelen (Member)

Dr. Kenneth D. Cashin (Member)

Dr. Hariharan Swaminathan (Member)

Dwight W. Allen
Dean
School of Education

December 17, 1973

Computer Simulation Study of Tailored Testing
Strategies for Objective-Based Instructional Programs

John P. Spinetti
University of Massachusetts

Abstract

Among the novel approaches to monitoring student progress in objective-based instructional programs is the technique of tailored testing. Tailored testing has been defined as a strategy for testing in which the sequence of test items a student receives is dependent on his performance on earlier items.

The amount of testing required in objective-based programs has been criticized, but to some extent this amount can be justified on the grounds that testing is an integral part of the instructional process. Nevertheless research is needed on strategies that offer the potential for reducing testing time that do not result in any appreciable loss in the quality of decision making from test results. In objective-based programs where the objectives can be arranged into hierarchies, one promising strategy is that of tailored testing. Some excellent work has been done by Richard Ferguson but his work provides only a beginning to the rather complicated problem of developing efficient tailored testing strategies. In addition, because of the complexity of his procedures, his tailored testing strategies can be applied only to computer-administered tests. Unfortunately many objective-based programs at the present time are implemented without the aid of computer terminals for instruction and testing. Therefore, what is needed is data pertaining to the effectiveness of a multitude of tailored testing

strategies, strategies that can be implemented without the aid of computers, to facilitate the eventual implementation of tailored testing.

This study was designed to investigate the effects of several factors on the quality of decision making and on the amount of testing time in a tailored testing situation. The factors chosen for study because of their perceived importance were: Test length, starting point, routing strategy, ability distribution, hierarchical structure of the objectives, and observed mastery cutting score.

In view of the complexity of conducting empirical studies to investigate the effectiveness of a large number of tailored testing strategies it seemed rather clear that a computer simulation approach was the best way to proceed. It was felt that such an approach would produce some interesting results on the relative value of different tailored testing strategies. These results would then provide a basis for designing some follow-up empirical studies on carefully selected levels of the more interesting factors as revealed by the simulation study.

A tailored testing strategy was defined by a test length, a mastery cutting score, a starting point, and a routing method. The two criterion measures for evaluating the tailored testing strategies were the number of objectives tested and the agreement between true and observed mastery records.

The results indicated that it was possible to obtain an overall reduction of more than 50% in testing time by utilizing a tailored testing procedure as an alternative to a conventional testing procedure.

Furthermore among the various tailored testing strategies, testing time was lower by about 15% when testing was initiated in the middle rather than at the extremes of the learning hierarchy. With regard to the number of errors of classification, as expected, the longer the test the fewer the average number of errors. There was, on the average, a 40% reduction in the number of errors of classification with a five-item test and an observed mastery cutting score of 0.80 when compared to a one-item test with an observed mastery cutting score of 1.00. Also, it was noted that the positioning of the cutting score affected the type of classification errors. If the observed cutting score was set above the true mastery cutting score, the majority of errors were of the false-negative type; however, if it was set lower, the highest percentage of errors were of the false-positive type. The results of the study were basically the same for two different learning hierarchies although tailored testing strategies with the hierarchy with two terminal objectives were slightly less favorable than with the hierarchy with a single terminal objective. Finally, although the analysis of routing methods was not extensive, this work did indicate the importance of routing methods on the overall effectiveness of a tailored testing strategy.

ACKNOWLEDGMENTS

I would like to extend my appreciation to the following people in recognition of the assistance and encouragement that they have given me throughout the course of this study.

Dr. Ronald K. Hambleton, who from my first contact with test theory and later as my major dissertation advisor, served to stimulate ideas and to provide encouragement. His guidance and assistance will always be remembered.

Dr. Leverage J. Thelen, who as my advisor in science education, provided my initial interest in research in science education.

Dr. Kenneth D. Cashin, who during my early years in science, influenced my appreciation of education as a field of study.

Dr. Hariharan Swaminathan, who during the writing of this study, provided valuable assistance and suggestions.

Deloris Fremouw, whose persistent dedication in assisting in the tabulation and drafting of this study is gratefully appreciated.

Special acknowledgment goes to my wife, Linda, whose patience, understanding, and encouragement played such an important role during the course of this study.

I would like to express my appreciation for the contribution of my parents for the advice and assistance they have provided me over the many years of my formal education. My mother's encouragement and sincere interest in my study will long be remembered. Finally this dissertation is dedicated to the memory of my father whose insight into problem solving served to develop my first interest in research.

TABLE OF CONTENTS

Abstract.....	ii
Acknowledgments.....	v
List of Tables.....	viii
List of Illustrations.....	xi
 Chapter I - Introduction.....	 1
1.1 Background.....	1
1.2 Objective-Based Instruction, Testing and Measurement.....	2
1.3 Statement of the Problem.....	4
1.4 Purposes of the Investigation.....	5
 Chapter II - Review of the Literature.....	 8
2.1 Introduction.....	8
2.2 Tailored Testing and Measurement Problems.....	8
2.3 Tailored Testing and Classification Problems.....	13
2.4 Summary.....	18
 Chapter III - Methodology.....	 20
3.1 Introduction.....	20
3.2 Terminology.....	20
3.3 Discussion of the Factors Under Consideration.....	22
3.4 Discussion of the Fixed Factors.....	31
3.5 Criterion Measures.....	36
3.6 Procedure.....	36
 Chapter IV - Results and Discussion.....	 40
4.1 Introduction.....	40
4.2 A Comparison of Tailored Testing and Conventional Testing...	40
4.3 Investigation of the Importance of the Factors of Starting Point and Ability Distribution on the Number of Errors of Classification.....	50
4.4 Investigation of the Importance of the Factors of Starting Point and Ability Distribution on the Number of Objectives Tested.....	52
4.5 Investigation of the Importance of the Factors of Test Length, Cutting Score, and Ability Distribution on the Number of Errors of Classification.....	52

TABLE OF CONTENTS (continued)

4.6	Investigation of the Importance of the Factors of Test Length, Cutting Score, and Ability Distribution on the Number of Objectives Tested.....	55
4.7	Investigation of the Importance of the Factors of Test Length, Cutting Score, and Ability Distribution on the Percentage of Type I Errors of Classification.....	55
4.8	A Comparison of the Effectiveness of Three Routing Procedures.....	58
Chapter V - Conclusions.....		62
5.1	Concluding Remarks.....	62
5.2	Suggestions for Further Research.....	65
References.....		67

LIST OF TABLES

Table	Page
3.3.1 A List of Linear Sequences for Gagné's Hydrolysis of Salts Hierarchy.....	29
3.3.2 A List of Linear Sequences for Ferguson's Addition-Subtraction Hierarchy.....	29
3.4.1 Summary Statistics for Various Tailored Testing and Conventional Testing Simulations with Hierarchy A (Totals Based on a Sample of 50 Examinees).....	32
3.4.2 Summary Statistics for Various Tailored Testing and Conventional Testing Simulations with Hierarchy A (Totals Based on a Sample of 100 Examinees).....	33
3.4.3 Summary Statistics for Various Tailored Testing and Conventional Testing Simulations with Hierarchy A (Totals Based on a Sample of 200 Examinees).....	34
3.4.4 Summary Statistics for Various Tailored Testing and Conventional Testing Simulations with Hierarchy A (Totals Based on a Sample of 500 Examinees).....	35
4.1.1 Summary Statistics for Various Tailored Testing and Conventional Testing Simulations with Hierarchy A (Totals Based on a Sample of 100 Examinees).....	41
4.1.2 Summary Statistics for Various Tailored Testing and Conventional Testing Simulations with Hierarchy B (Totals Based on a Sample of 100 Examinees).....	45
4.3.1 Summary of the Average Number of Errors of Classification for Combinations of Levels of Ability Distribution and Starting Point Factors with Hierarchy A (Totals Based on a Sample of 100 Examinees).....	51
4.3.2 Summary of the Average Number of Errors of Classification for Combinations of Levels of Ability Distribution and Starting Point Factors with Hierarchy B (Totals Based on a Sample of 100 Examinees).....	51
4.4.1 Summary of the Average Number of Objectives Tested for Combinations of Levels of Ability Distribution and Starting Point Factors with Hierarchy A (Totals Based on a Sample of 100 Examinees).....	53

LIST OF TABLES (continued)

Table	Page
4.4.2 Summary of the Average Number of Objectives Tested for Combinations of Levels of Ability Distribution and Starting Point Factors with Hierarchy B (Totals Based on a Sample of 100 Examinees).....	53
4.5.1 Summary of the Average Number of Errors of Classification for Combinations of Levels of Ability Distribution, Cutting Score and Test Length Factors with Hierarchy A (Totals Based on a Sample of 100 Examinees).....	54
4.5.2 Summary of the Average Number of Errors of Classification for Combinations of Levels of Ability Distribution, Cutting Score and Test Length Factors with Hierarchy B (Totals Based on a Sample of 100 Examinees).....	54
4.6.1 Summary of the Average Number of Objectives Tested for Combinations of Levels of Ability Distribution, Cutting Score and Test Length Factors with Hierarchy A (Totals Based on a Sample of 100 Examinees).....	56
4.6.2 Summary of the Average Number of Objectives Tested for Combinations of Levels of Ability Distribution, Cutting Score and Test Length Factors with Hierarchy B (Totals Based on a Sample of 100 Examinees).....	56
4.7.1 Summary of the Average Percent of Type I Errors of Classification for Combinations of Levels of Ability Distribution, Cutting Score and Test Length Factors with Hierarchy A.....	57
4.7.2 Summary of the Average Percent of Type I Errors of Classification for Combinations of Levels of Ability Distribution, Cutting Score and Test Length Factors with Hierarchy B.....	57
4.8.1 Summary of the Average Number of Errors of Classification for Three Levels of the Ability Distribution Factor and Three Routing Strategies with Hierarchy A (Totals Based on a Sample of 100 Examinees).....	59

LIST OF TABLES (continued)

Table	Page
4.8.2 Summary of the Average Number of Errors of Classification for Three Levels of the Ability Distribution Factor and Three Routing Strategies with Hierarchy B (Totals Based on a Sample of 100 Examinees).....	59
4.8.3 Summary of the Average Number of Objectives Tested for Three Levels of the Ability Distribution Factor and Three Routing Strategies with Hierarchy A (Totals Based on a Sample of 100 Examinees).....	60
4.8.4 Summary of the Average Number of Objectives Tested for Three Levels of the Ability Distribution Factor and Three Routing Strategies with Hierarchy B (Totals Based on a Sample of 100 Examinees).....	60

LIST OF ILLUSTRATIONS

Figure	Page
3.3.1 Gagné's Hydrolysis of Salts Hierarchy.....	27
3.3.2 Ferguson's Addition-Subtraction Hierarchy.....	28

Chapter I

Introduction

1.1 Background

The overall goal of an individualized instructional program is to provide an educational program which is maximally adaptive to the requirements of the individual learner. Instructional objectives generally provide the structure for the curriculum and serve as a basis for the development of lesson materials and tests. A basic assumption of many individualized programs is that students can progress most effectively if they proceed through a sequence of structured objectives at a rate which will allow the student to master each objective before proceeding on to the next.

While suggestions for developing individualized instructional programs were made as early as 1922 by Washburne (1922) and more recently by Wilhelms (1962), it has only been in the last few years that we have seen any implementation of large-scale individualized instructional programs. Among the better-known large-scale programs are: Individually Prescribed Instruction (IPI) (Glaser, 1968, 1970), Program for Learning in Accordance with Needs (PLAN) (Flanagan, 1967, 1969), Computer-Assisted Instruction (CAI) (Suppes, 1966; Atkinson, 1968; Atkinson and Wilson, 1969), Individualized Mathematics Curriculum Project (DeVault, Kriewall, Buchanan, and Quilling, 1969), and Mastery Learning (Carroll, 1963, 1970; Bloom, 1968, and Block, 1971).

At the present time there is little information available on testing methods and decision making procedures for individualized instructional programs. In part this is because of the rather different nature

of the testing problems encountered in these new programs. Needed are criterion-referenced tests which are designed to measure student performance on specific behavioral objectives along an absolute ability continuum. These tests differ from norm-referenced tests which are principally designed to facilitate making comparisons among students (Popham and Husek, 1969; Glaser and Nitko, 1971; Hambleton and Novick, 1973).

Among the novel approaches to monitoring student progress via the use of criterion-referenced tests in individualized instructional programs is the technique of branched, sequential, or tailored testing (Lor, 1970; Ferguson, 1969). Tailored testing has been defined as a strategy for testing in which the sequence of test items a student receives is dependent on his performance on earlier items.

1.2 Objective-Based Instruction, Testing and Measurement

Among the basic features of individualized instructional programs as developed by Glaser (1970) and Glaser and Nitko (1971) are two components, one which relates to the goals of learning which are stated in behavioral terms and a second which relates to the assessment of student performance which is continuously monitored. In certain programs the goals of learning are represented in a hierarchical structure. Objectives within the hierarchy may be arranged in a series of linear sequences. In testing objectives on such a structure one can make inferences about objectives in the hierarchy which have not been tested. If, for example, an examinee is found to have proficiency in a specified

objective, all objectives prerequisite to it are also considered mastered. If the examinee lacks proficiency in an objective it can be inferred that all objectives to which it is prerequisite are also unmastered.

The primary measurement problem in most individualized instructional programs, at least at this stage of their development, is one of determining if a student's true mastery level, π_i , is greater than or equal to a specified standard, π_o . The value π_i is the "true" score for an individual "i" in the domain of items measuring the behavioral objective where mastery assessment is of interest. Since, in theory at least, there are an unlimited number of items in the domain, a small sample of items is generally selected from the domain to obtain an estimate of the student's true mastery level, represented by $\hat{\pi}_i$. The standard π_o is the cutting score used to partition individuals into the mastery and non-mastery states on particular objectives in the hierarchical structure.

In order to design instructional programs so as to provide instruction which is responsive to individual differences, continuous and reliable feedback of information is necessary to place each student at a point in the curriculum commensurate with his performance level. An example of a program where this is done is the Individually Prescribed Instruction (IPI) program. For example, in this program the curriculum is defined in terms of behavioral objectives which are organized into areas, levels, and units. In such a program the test constructor must provide instruments to facilitate the placement and diagnosis of pupil achievement.

Lindvall and Cox (1970) in a study of evaluation of the IPI program

state:

Under this plan for individualization each time a pupil faces a new unit of instruction during any part of the school year, the teacher must have information about what that pupil does or does not know in order to plan an efficient and effective program tailored to the individual. As the pupil proceeds by working on the appropriate instructional materials, the teacher monitors the pupil's progress in mastering the behavioral skills in his particular instructional sequence, so that an efficient instructional program can again be planned to that pupil. A continuous diagnosis of performance provides necessary feedback for long-range planning and periodic modification, when necessary, of individual programs. Diagnosis of pupil performance thus allows for pupil placement in the instructional sequence and provides for a continuous evaluation of pupil progression throughout the course of study.

To assist in placing a student at the appropriate place in the curriculum it has been suggested that a tailored testing strategy may be helpful. Tailored testing strategies have been shown to significantly reduce the amount of testing time without any loss of accuracy in classifying students into mastery states (Ferguson, 1969).

1.3 Statement of the Problem

There are many problems associated with the development and implementation of objective-based programs. Among the problems are those related to testing. The amount of testing required in objective-based programs has been criticized, but to some extent this amount can be justified on the grounds that testing is an integral part of the instructional process. Nevertheless research needs to be done on strategies that offer the potential for reducing testing time that do not result in any appreciable loss in the quality of decision making from test results. In objective-based programs where the objectives

can be arranged into hierarchies, one promising strategy is that of tailored testing (Ferguson, 1969). Some excellent work was done by Ferguson but his work provides only a beginning to the rather complicated problem of developing efficient tailored testing strategies. In addition, because of the complexity of his procedures, his tailored testing strategies can be applied only to computer-administered tests. Unfortunately many objective-based programs at the present time are implemented without the aid of computer terminals for instruction and testing. Therefore, what is needed are data pertaining to the effectiveness of a multitude of tailored testing strategies, strategies that can be implemented without the aid of computers, to facilitate the eventual implementation of tailored testing.

1.4 Purposes of the Investigation

Tailored testing is a technique or strategy that has been suggested for use in objective-based programs because of its potential for reducing the amount of testing time without any loss in the quality of decision-making via the use of criterion-referenced test data.

This study was designed to investigate the effects of several factors on the quality of decision making and on the amount of testing time in a tailored testing situation. The factors chosen for study because of their perceived importance were: test length, starting point, routing strategy, ability distribution, hierarchical structure of the objectives, and observed mastery cutting score.

Rather obviously the number of items (test length) used to assess mastery of objectives will be a factor in determining the overall effective-

ness of a tailored testing strategy - other things being equal, the longer the test the more accurate will be the decision-making. Whether one chooses to initiate testing at the bottom, middle or top of the hierarchy and how one chooses to route a student from one objective to another in the hierarchy will also affect the efficiency of a tailored testing strategy. The level of ability of the examinees is a factor also, since for example, a tailored testing strategy which utilizes prior information on the abilities of the examinees will be better than one that does not. Finally, the effectiveness of any tailored testing strategy is likely to be dependent on the particular hierarchical structure of the objectives.

An important consideration in our work was that tailored testing strategies under investigation be implementable without the aid of computer terminals. The primary effect of this was that it ruled out the possibility of using complex decision-making rules such as the one adopted by Ferguson (1969). Also in our work we required a fixed number of items to assess mastery of each objective tested, items were scored right or wrong, and all items measuring a particular objective were assumed to have similar statistical properties. Individual performance on the test items was assumed to be represented by the binomial test model (Lord and Novick, 1968). Let us be quick to point out however that we certainly do not believe that the tailored testing strategies we studied could be implemented easily. Much research would still need to be done on the implementation phase. Our primary purpose was to produce some data on the effectiveness of various tailored testing strategies that could be implemented in an objective-based program and

that would not require a computer terminal for the administration of the test items.

In view of the complexity of conducting empirical studies to investigate the effectiveness of a large number of tailored testing strategies it seemed rather clear that a computer simulation approach was the best way to proceed. It was felt that such an approach would produce some interesting results on the relative value of different tailored testing strategies. These results would then provide a basis for designing some follow-up empirical studies on carefully selected levels of the more interesting factors as revealed by our simulation study. One benefit of adopting a computer simulation approach was that we eliminated some of the problems that are related to item validation (Rovinelli and Hambleton, 1973), item selection (Glaser and Nitko, 1971), and establishment of the hierarchical structure of a set of objectives.

Chapter II

Review of the Literature

2.1 Introduction

Work on tailored testing has only recently attracted the attention of educational researchers. While there were several studies in the 1950s and early 1960s, Lord's recent work in the area (Lord, 1970; Lord, 1971 a, b, c, d, and e) has done much to bring attention to it. Lord has mainly been concerned with improving the precision of measurement of an examinee's latent ability and at the same time, decreasing the amount of testing time.

Ferguson's work in 1969 typifies a second line of research in tailored testing. It is an adaptation of tailored testing to situations in which the testing problem is one of classifying individuals into mastery states rather than precisely estimating their ability.

In the remainder of this chapter we have provided a brief review of both lines of research. The chapter is concluded with a summary section.

2.2 Tailored Testing and Measurement Problems

According to Lord (1970) a tailored test is a test in which each item is selected on the basis of the examinee's responses to previous items. The objective is to match the item difficulty to the examinee's ability. Early testing of this type has been referred to as "branched," "programmed," "sequential item," or "computerized" testing. The key concept is that an examinee's ability is measured most effectively when the test items are neither too difficult nor too easy for him. The value

of tailored testing is therefore primarily for those examinees for whom the conventional test would be too difficult or too easy; however, considerable savings in testing time can be effected for all examinees as will be demonstrated later in this review.

In the area of testing for the purpose of measurement a number of studies have been performed. The procedure as described by Lord (1970) is concerned with "measuring" a single examinee's ability with respect to one psychological dimension. The problem then becomes one of estimating or determining the ability level of the examinee on a psychological continuum using n test items as efficiently as possible. An optimum strategy for item selection is unknown since the examinee's ability level is unknown, but a "best" strategy has been tried by Lord. A form of stepping or branching is applied whereby if an examinee's true score is equal to or above the cutting score, the next item administered is more difficult, and if his true score is below the cutting score, a less difficult item is administered. A shrinking-step-size procedure in which a large step size in terms of item difficulty is applied for the first few items so that the item difficulty can be rapidly adjusted to suit an extreme examinee is suggested by Lord. In this strategy smaller step sizes are ultimately applied to converge on the "best" value of the examinee's ability.

Lord also suggests two strategies for reducing the effects of chance success. To insure that the examinee will eventually be answering 60 to 70 percent of the items correctly, he recommends that the step size in the positive direction be made smaller than the step size in the negative direction. Alternate to this procedure or as a supplement Lord also suggests that items of equal difficulty could be

combined into blocks. The score on each block of items could be used to determine which subsequent block of items to administer to the examinee next. When this technique is applied, the effect of guessing correctly on one item does not have as great an influence on the branching procedure.

Cleary, Linn and Rock (1968a) performed an exploratory study of tailored testing with the objective of reducing testing time in large scale testing and of producing information for evaluation purposes. The simulated tests consisted of two sections: (1) a routing section of 20 items with branching to direct the examinee to appropriate items and (2) a measurement section. In this study it was found that sequential item sampling based on a point biserial correlation with the total test score was the best routing procedure since it resulted in the fewest errors in classification and the highest correlation with total test score. However the sequential item sampling technique was also the most difficult method to employ. Cleary suggests obtaining some predictors of optimal item selection.

In a follow-up study by Cleary, Linn and Rock (1968b) the simulated reproduction of total test score through the use of tailored tests was investigated. Two variations on the test strategy were used: (1) the items were scored sequentially and (2) examinees were assigned to a group along a total continuum of ability. The procedure resulted in high correlations with total test score. It was suggested that the best items for routing might not be the best items for measurement. In addition it was recommended that the validity of tailored testing be

investigated under actual test administration. It should be noted that both of Cleary's studies used a single internal criterion.

Linn, Rock and Cleary (1969) investigated the use of external criteria in the development and evaluation of several tailored testing methods. A year and a half following Cleary's studies, scores were obtained on the College Board Achievement Test for about two-thirds of the original population. Five routing strategies and two branching strategies were utilized. The five routing strategies were the same as those used by Cleary. The two branching methods were based on a complete branched tree design as compared to a routing section. One was a ten-item tailored test with branching after each item according to whether a correct or incorrect response was obtained. The other was an item-block tailored test where branching occurred after each block of five items. The results of the tailored tests were compared with shortened conventional tests of 10, 20, 30, 40, and 50 items. The comparison was based on the equivalent length of a test which would be parallel to a 190-item test. The best results for this simulation were obtained from a tailored testing method.

Two tailored testing procedures for dichotomous decisions were investigated by Linn, Rock and Cleary (1970) using data of 4080 college students. The purpose of the study was to investigate the potential advantages of a tailored testing procedure for multidimensional dichotomous designs. The 4080 students were divided into two equal numbered groups--2040 students for short tailored testing procedures and short conventional tests and 2040 students for cross-validation purposes. In this study existing item response data on three achievement tests were

used to compare the use of two tailored testing procedures with conventional testing for purposes of obtaining dichotomous decisions on each of three dimensions. The three achievement tests used were in English composition, mathematics, and natural sciences. Two procedures were employed. In the first each dimension was treated independently. In the second the decision in mathematics was utilized in a decision for English composition and for natural sciences. The results showed that for a given degree of accuracy the first tailored testing method required one-half as many items as a short conventional test. The efficacy of the second tailored testing method using knowledge on one dimension in making the assignment on the second dimension depended upon the relationship between the two dimensions.

Lord (1971c) described a study involving a self-scoring flexilevel tailored test in which the examinee starts with a "middle" difficulty item and is routed to items depending upon whether or not he answers the item correctly. The prerequisite for such an application requires the unidimensionality of the items in the test. Lord suggests that 2-stage testing could be utilized in relaxing the unidimensionality of the items, but this approach would require complicated scoring procedures. In Lord's flexilevel test, although different sets of items are taken by different examinees, scoring methods provide comparable scores for all under a self-scoring format.

In a study by Stocking (1971) the effect of test length was investigated for tailored tests with a variety of designs and scoring procedures. The two methods of design involved the fixed-step-size and the shrinking-step-size procedure. The scoring procedures were the

final difficulty score and the average difficulty score. It was concluded that the "best" procedure would be the shrinking-step-size procedure with a final difficulty score, but this procedure would be difficult to administer since it requires a very large pool of items. The fixed-step-size procedure using an average difficulty score provided nearly as good a measurement as other methods which used shrinking-step-size while requiring far fewer items.

In a theoretical study of two-stage tailored testing, Lord (1971b) concluded that if items cannot be answered by guessing, two-stage testing procedures are about as effective as up-and-down procedures including the fixed-step-size procedure. However if guessing "works" 20 percent of the time, no two-stage procedure matched the "best" up-and-down procedure. As an extension of his study Lord suggested the use of more second stages--ten or more--or the use of a three stage procedure.

2.3 Tailored Testing and Classification Problems

The studies reviewed in the previous section were concerned with the problem of measurement--estimating a person's standing on a latent trait from his responses to a series of ability test items. In some of the earliest studies of using tailored testing for decision-making, Cowden (1946) applied Wald's (1947) sequential testing for assigning grades in a statistics class. Cowden's study used subsets of 20 items administered one at a time selected from a pool of 200 items. Within certain pre-specified error tolerances, he found that a decision could be made with most students by using less than one-third of the items

available.

Using real data simulation techniques to make a dichotomous pass/fail decision, Moonan (1950) applied the same tailored testing methods. His results showed that an average of 40 items was necessary to approximate the decision which could be made on the basis of 75 items.

Cronbach (1966) as early as 1954 and Cronbach and Gleser (1965) suggested the application of tailored testing item presentation procedures for categorical decision-making. Further support was contributed in a theoretical finding by Green (1970). He reported a minimum savings of 50 percent in testing time in a study involving a tailored testing procedure. The items selected for this study were of equal discriminating power and were assumed free of chance success due to guessing.

Green (1970), in response to a paper by Lord (1970), indicated that although tailored testing does not always improve upon conventional testing in its application to measurement, it could serve a useful function in decision-making. Green states:

Lord's gloomy conclusions about tailored testing arise from considering measurement per se rather than any use to which the measures are put. This restricted outlook is in tune with our current wasteful decision-making procedures in education, industry, and the military establishment. We typically measure first and decide later. There is very seldom any interplay between measurement and decision. No allowance is made for a decision to collect more data. Cronbach and Gleser (1965) have argued forcibly that this procedure wastes information and ignores many aspects of the decision problem. Both selection and placement problems, they argued, need to be integrated with the measurement process in an overall decision procedure that includes a consideration of costs and values. Although many measurement experts, including Lord (1962), are sympathetic with Cronbach and Gleser's thesis, practical matters have so far interfered with

any serious attempt at implementation. Until now it has been impractical to obtain some measures, consider their implications, and then possibly obtain further measures. The computer-controlled test overcomes the practical difficulties, and serious thought can now be given to an integrated decision system.

In emphasizing the role of tailored testing in individualized instructional programs, Green indicated that when a test is to be used to determine placement of students on a hierarchy of skills, a tailored test would be preferred to a conventional test since each student would be asked to answer only those items which provide information on an individual at his level of proficiency. The major advantage of computerized tailored testing is in its ability to stop the testing procedure as soon as a decision can be made. The role of the testing procedure is therefore to tailor the number of test items required for a decision and not their difficulty.

Linn et al. (1969) suggested that there are some advantages beyond the psychometric characteristics of the test that should be looked at. Although the precision of measurement may not be improved, by minimizing the number of items which are too easy or too difficult for an individual, the result could be a more challenging and interesting test for the examinee. Furthermore, the possibility exists that testing time could be reduced without a reduction in precision of measurement. These considerations are particularly important in an individualized instructional program where continuous monitoring of student progress through testing consumes a large fraction of the time that a student spends in a program.

There are relatively few studies reported in the area of tailored testing for classificatory purposes but one by Ferguson (1969) using a

hierarchical structure of skills in the University of Pittsburgh's Individually Prescribed Instruction Program is of particular importance. In his study he was concerned with classifying students with respect to mastery or non-mastery at each level of a hierarchical structure and thereby locating an individual's level of proficiency on the learning hierarchy. To accomplish this, computer-based tailored testing was applied to a hierarchy of skills in an objective-based curriculum.

In the testing procedure, sequences of objectives can be selected from the hierarchy and can be listed numerically. Objectives listed numerically lower in a sequence are prerequisite to those higher in the sequence. Tailored testing on a hierarchical structure involves selecting a starting point--for example, a middle difficulty objective. The number of items administered to assess mastery of an objective is referred to as the test length. A sufficiently small number of questions is chosen so as to minimize testing time while not adversely affecting test reliability and test validity. Following the administration of each item a sequential probability ratio test is used to classify each student into one of three categories: 1) mastery; 2) non-mastery; or 3) no decision. When "no decision" occurred an additional item was administered and the probability ratios were re-calculated. For each objective, items were administered until the student had a score p_0 or above, in which case he was said to have mastery, or until the student had a score p_1 or below, wherein he was classified as non-mastery. According to Ferguson (1969), " p_0 and p_1 are specified such that indicating sufficient proficiency in the objective is an error of grave consequence only if

$p \leq p_1$ [where p is the proportion correct] and indicating insufficient proficiency in the objective is a serious error only if $p \geq p_0$." In Ferguson's model p_0 does not equal p_1 and no decision for routing is made while the student's score lies between p_0 and p_1 . Since the decision for mastery or non-mastery was not always made on the basis of a small number of questions, the items administered could under certain circumstances be extended to a fairly high number. If no decision were made after thirty questions, the testing was terminated at that point. The decision under this circumstance was based upon whether the student's score was closer to mastery or non-mastery. The routing strategy that Ferguson used was fairly complex and a computer was necessary to perform the actual routing.

The final phase of Ferguson's study consisted of administering his computerized tailored testing, classification system to 75 students in grades 1 to 6 and comparing the results with a paper and pencil conventional test administration. He found a 60 percent savings in time in the computerized administration using a variety of branched test models. A test-retest of the tailored testing procedure gave high reliability, with the reliabilities of the tailored testing classifications higher than those of the paper and pencil conventional test approach. The validity of the tailored testing approach was also found to be high.

In addition, for purposes of placing students within the curriculum, the complete routing model could be used while intensive testing within a unit of work could supplement the complete routing model

for specific placement within each unit. Ferguson indicated that the study of tailored testing as applied to Individually Prescribed Instruction should prove most interesting from a theoretical point of view. He indicated that tailored testing in the IPI program, where extensive testing is required, is ideal for investigative research and could result in a reduction of testing time while providing test information at least comparable to conventional testing.

2.4 Summary

The review of the literature indicates that tailored testing appears to have a high potential in its application to testing in individualized instructional programs involving a hierarchy of objectives. It is clear from the research reported that there is a vast array of tailored testing procedures which have been proposed, and undoubtedly still many more to be suggested. However in the area of making instructional decisions only one study involving an empirical application has been reported.

Research available on tailored testing shows considerable promise for the superiority of these methods over conventional testing procedures. Using a variety of research approaches and a number of different criteria, tailored tests have been shown to be: 1) considerably shorter than conventional tests, with little or no loss in validity or reliability; 2) more reliable than conventional tests and yielding more nearly constant precision than standard tests over a wide range of abilities; and 3) in many cases more valid than are conventional tests. Tailored tests also have the potential of being more fair to extreme group members in that they are less likely to result

in frustrating experiences, thereby resulting in less errors of measurement.

Chapter III

Methodology

3.1 Introduction

The study was designed to investigate the interactive effects of six factors assumed to be important in tailored testing research. The six factors were: test length, starting point, routing strategy, observed mastery cutting score, hierarchical structure of the objectives and ability distribution of the examinees. The criteria chosen to evaluate the effectiveness of a tailored testing strategy were the accuracy of classification decisions relating to mastery and the amount of testing time. Because of the complexity of the problem, an analytic solution was not possible; hence it was decided to conduct the investigation using a computer simulation approach. The remainder of the chapter outlines the methodology for our investigation.

3.2 Terminology

In order to facilitate our discussion of tailored testing research it was necessary to define several terms. When we refer to test length, we mean the number of test items that are used to measure the mastery of a single behavioral objective. The true mastery score for an individual denoted π is defined as the proportion of items in the domain of items measuring a behavioral objective that the individual can answer correctly. An individual has a true mastery score for each objective in the learning hierarchy. Typically a cut-off score or threshold score called in our terminology, the true mastery cutting score, denoted π_0 , is set to separate individuals into one of two true mastery states. (In general, the true mastery cutting score varies from one objective

to another but in our work it was set at a constant value of 0.8.) An individual is classified as a true master if $\pi \geq \pi_0$, and as a true non-master if $\pi < \pi_0$. The information on an individual's true mastery states across the objectives in a learning hierarchy is defined as his true mastery record.

Since all items in the domain of items defined by a behavioral objective cannot usually be administered to an individual for the purpose of assessing his true mastery score, some small number of items are sampled. The observed mastery score¹, $\hat{\pi}$, is defined as the proportion of items that an individual answers correctly of the items sampled. The observed mastery cutting score, $\hat{\pi}_0$, is defined as the proportion of items on the test measuring an objective that the instructional designer deems necessary for the individual to answer correctly to be classified as a master. An individual is classified as a master if $\hat{\pi} \geq \hat{\pi}_0$, and as a non-master if $\hat{\pi} < \hat{\pi}_0$. The information on an individual's observed mastery states across the objectives in a learning hierarchy is defined as his observed mastery record.

Conforming with the notation employed by Hambleton and Novick (1973) we defined a two-valued parameter ω to denote the mastery state of an individual on a particular behavioral objective. If the individual is a true non-master, i.e., if $\pi < \pi_0$, we set, $\omega = 0$, and if he is a master, i.e., $\pi \geq \pi_0$, $\omega = 1$.

Now if we estimate π from an individual's performance on a sample of test items, then an estimate of ω can be obtained in the following way:

¹We note that the observed mastery cutting score is usually set by the instructional designer and is not estimated from the data as the symbol used for the cutting score might suggest.

$$\hat{\omega} = 1, \text{ if } \hat{\pi} \geq \hat{\pi}_0 \text{ and}$$

$$\hat{\omega} = 0, \text{ if } \hat{\pi} < \hat{\pi}_0.$$

Note that $\hat{\pi}_0$ is usually set as close to π_0 as possible.

It is now possible to think of an individual's true mastery record and observed mastery record as collections of 1's and 0's denoting mastery and non-mastery states across the objectives in the learning hierarchy. It is clear that the error of classification of an objective, $e = (\hat{\omega} - \omega)$, takes on one of three values, +1, -1, or 0 depending on whether we make a false-positive error (Type II error), false-negative error (Type I error)¹, or no error in classification. A false-positive error occurs when an individual is estimated to be a master when he is not (i.e., $\hat{\pi} \geq \hat{\pi}_0$ and $\pi < \pi_0$). A false-negative error occurs when an individual is estimated to be a non-master when the reverse is true ($\hat{\pi} < \hat{\pi}_0$ and $\pi \geq \pi_0$).

3.3 Discussion of the Factors Under Consideration

(a) Test Length

Test length is one of the most important variables in designing a tailored testing strategy. In our investigation tests of length 1, 2, 3, 4, and 5 items were considered. This allowed us to span the range of test lengths typically used by developers of objective-based programs (Hambleton, 1973).

(b) Starting Point

Of course there are as many starting points for a tailored testing strategy as there are objectives in the learning hierarchy. For practical

¹The usage of the terms, Type I and II errors, should not be confused with their usage in hypothesis testing.

reasons we studied only four starting points from each hierarchy of objectives but they were strategically located to be representative of available starting points at the bottom, middle and top of the hierarchy.

(c) Routing Strategies

Of the many possible routing strategies, a decision was made to use a simple fixed-step-size procedure for the majority of our analyses. In this study that meant that when an individual demonstrated mastery (non-mastery) of a particular objective $\hat{\pi} \geq \hat{\pi}_0$ ($\hat{\pi} < \hat{\pi}_0$), he was moved to the next highest (lowest) objective in the linear sequence of objectives being tested. Two additional variable-step-size procedures were also explored. The first method involved routing the examinee to the top (bottom) of the learning hierarchy if he was estimated to be in a mastery state (non-mastery state) on the initial objective tested. Subsequent routing was based on the simple fixed-step-size procedure. The second method involved routing an examinee to the top (bottom) of the learning hierarchy if he demonstrated 100% (0%) mastery of the initial objective tested in the learning sequence. The usual simple fixed-step-size procedure was used in cases where examinees did not achieve one of the two extreme mastery scores. Subsequent routing was also based on the simple fixed-step-procedure.

Of course there exists an extremely large number of additional routing procedures that could have been tried. Complex ones were eliminated from our investigation since they would be difficult, if not impossible, to operationalize in a classroom testing situation. Only

three simple ones were tried, in part, this was for economic reasons, and also because it was our feeling that three routing strategies were sufficient to provide for some interesting comparisons.

Let us look now in more detail at the three routing methods. To discuss the routing methods it is convenient to think of the linear sequences of each hierarchy arranged into a matrix form. The columns of the matrix correspond to the linear sequences and the rows to the levels of difficulty of objectives in the sequences. The entries of the matrix are the numbers of the objectives shown in Figures 3.3.1 and 3.3.2. This meant that with Gagné's hierarchy we had a matrix of six rows and eight columns. With Ferguson's hierarchy we had a matrix of nine rows and seven columns. An objective number appeared as many times in the matrix as there were linear sequences of which it was a part.

In simulating an individual's test performance through a tailored testing strategy with a fixed-step-size routing procedure the first step was to estimate his mastery state on the objective selected as the starting point. (Details on how this was done are described in Section 3.6. In this section we will concentrate our attention only on the routing methods.) If the individual is estimated to have mastered the objective, he is given credit for having mastered all prerequisite objectives and no testing is done on these objectives. He is then tested on the next highest objective in the sequence. In the opposite situation where an individual is estimated to be in a non-mastery state for the initial objective tested, he is assumed to be a non-master of all

objectives of which that objective is a prerequisite. He is then tested on the next lowest objective in the sequence. This process is continued until a mastery decision has been made for all objectives in the sequence and all related objectives in other sequences. Moving on to the next linear sequence the starting point for testing was taken to be the objective that was at the lowest level above the level of the last objective tested in the previous sequence, if the individual was estimated to be a non-master on the last objective tested in the previous sequence, then the starting point is the objective at the highest level below the level of the last tested objective. Once a starting point was determined the same process was repeated that was described for the first sequence. This process was repeated for each linear sequence of objectives in the hierarchy.

The second and third routing procedures were identical to the one just described except for the routing decision made on the basis of an individual's performance on the initial objective tested. The possible decisions have been described earlier. These last two routing procedures were similar in approach to suggestions made by Wood (1971) and Bayroff, Thomas, and Anderson (1960).

(d) Hierarchical Structure of the Objectives

Of the many learning hierarchies that are available in the literature two were selected for investigation. They were the learning structures for hydrolysis of salts (Gagné, 1965) and addition-subtraction (Ferguson, 1969). The second one was selected to facilitate comparison of some of our results with those of Ferguson. Flowcharts of the hierarchies are shown in Figures 3.3.1 and 3.3.2. A list of the linear sequences that make up each hierarchy are shown in Tables 3.3.1 and 3.3.2.

(e) Distribution of True Mastery Scores

The distribution of true mastery scores for individuals across the objectives in the two learning hierarchies was monitored by specifying the average level of performance and the distributional form of true mastery scores for the 100 examinees in a middle objective of each hierarchy. In Gagné's hierarchy (referred to as hierarchy A) the middle level objective was number 13 (see Figure 3.3.1). In Ferguson's hierarchy (referred to as hierarchy B) the middle level objective was number 13 (see Figure 3.3.2). It was assumed that true mastery scores in the middle level objective followed a beta distribution. First, a beta distribution rather than a normal distribution of true mastery scores was used since the beta distribution is defined on the same interval (0 to 1) as the true mastery scores whereas the normal distribution is not. More importantly it was relatively easy to generate skewed beta distributions of true mastery scores to approximate distributions that might be expected to occur with real test data. Also a normal distribution is a special case of the beta distribution.

Figure 3.3.1 Gagné's Hydrolysis of Salts Hierarchy

Figure 3.3.2 Ferguson's Addition-Subtraction Hierarchy

Table 3.3.1

A List of Linear Sequences for Gagné's
Hydrolysis of Salts Hierarchy

Sequence	Objectives Comprising the Sequence
1	1,6,10,12,15,16
2	2,7,10,12,15,16
3	3,8,11,13,15,16
4	3,8,11,14,15,16
5	4,9,11,13,15,16
6	4,9,11,14,15,16
7	5,9,11,13,15,16
8	5,9,11,14,15,16

Table 3.3.2

A List of Linear Sequences for Ferguson's
Addition-Subtraction Hierarchy

Sequence	Objectives Comprising the Sequence
1	1,4,7,10,12,13,14,16,17
2	1,9,13,14,16,17
3	1,6
4	2,6
5	3,6
6	2,5,8,11,15,18
7	3,11,15,18

The beta distribution is given by the expression

$$F(x) = \frac{(N - 1)!}{(R - 1)! (N - R)!} x^{R-1} (1 - x)^{N-R-1}.$$

The two parameters of the distribution, R and N, are related to the mean and variance of the distribution in this way,

$$E(x) = \frac{R}{N}, \quad \text{Var}(x) = \frac{R(N-R)}{N^2(N+1)}.$$

In our simulations, N was set equal to six and the values of R considered were two, four, and five, to generate positively skewed, negatively skewed and high negatively skewed distributions, respectively, of true mastery scores in a middle level objective. For particular values of N and R which specified the shape of the beta curve, and using numerical integration techniques, the mastery score scale (0, 1) was divided into ten unequal intervals where the end points of each interval were set in a way such that 10% of the area under the beta curve was included. In generating the initial true mastery scores for individuals on the middle level objective, 10% were selected at random, using a random number generator, along each of the ten intervals.

Two rules were used to monitor the generation of true mastery scores. The first was that in any linear sequence of objectives in the hierarchy the true mastery scores ranged from largest to smallest as one moved from the bottom to the top of the sequence. The second was that an individual was assigned true mastery scores that were roughly similar for objectives positioned at about the same general level of the hierarchy. (This

constraint was necessary to insure that we could satisfy the first rule.)

Beginning with the sequence in which the initial true mastery score for an individual was generated, the remaining were obtained in a straightforward manner. The true mastery score for the next highest objective in the sequence was generated at random from a uniform distribution on the interval $(0, \pi)$. The next lowest was generated on the interval $(\pi, 1)$. The process was repeated until true mastery scores were obtained for all objectives in the sequence. This process was repeated for the remaining sequence beginning in each sequence with the known true mastery score closest to the middle of the sequence and imposing the second rule mentioned above.

3.4 Discussion of the Fixed Factors

(a) True and Observed Mastery Cutting Scores

The true mastery score was set at a value of 0.8 for all of our analyses. This value has frequently been used to categorize individuals into dichotomous mastery states. The observed mastery cutting score was varied from one analysis to another depending on the test length. With the 1, 2, 3, 4, and 5 item tests it was set at 1.00. In addition, mastery cutting scores of .67, .75, and .80 were considered for the 3, 4, and 5 item tests, respectively.

(b) Selection of a Sample Size

Several computer simulations were carried out to determine the number of examinees required to produce stable estimates of the criterion measures that were used to describe the effectiveness of the various tailored testing strategies. The results of these simulations are summarized in Table 3.4.1 to 3.4.4 for samples of 50, 100, 200 and 500 examinees.

Table 3.4.1

Summary Statistics for Various Tailored Testing
and Conventional Testing Simulations with Hierarchy A
(Totals Based on a Sample of 50 Examinees)

Starting Point	Ability Distribution Parameters N R	Mastery Testing		Tailored Testing			No. of Objectives Tested	Conventional Testing		
		Length	No. Required For Mastery	Combined Errors	Type I Errors	Type II Errors		Combined Errors	Type I Errors	Type II Errors
13	6 2	1	1	232	16	216	580	228	50	178
13	6 2	2	2	238	34	204	632	214	76	138
13	6 2	3	2	160	6	154	560	226	14	212
13	6 2	3	3	156	82	74	706	194	132	62
13	6 2	4	3	174	4	170	578	156	28	128
13	6 2	4	4	134	104	20	734	182	158	24
13	6 2	5	4	94	10	84	630	162	42	120
13	6 2	5	5	146	130	16	734	204	184	20
13	6 4	1	1	194	10	184	504	220	46	174
13	6 4	2	2	140	46	94	550	204	130	74
13	6 4	3	2	204	4	200	506	226	20	206
13	6 4	3	3	116	60	56	608	196	142	54
13	6 4	4	3	150	10	140	504	148	24	124
13	6 4	4	4	122	74	48	618	200	162	38
13	6 4	5	4	132	16	116	542	178	42	136
13	6 4	5	5	128	114	14	662	192	158	34
13	6 5	1	1	136	20	116	470	162	48	114
13	6 5	2	2	116	40	76	512	114	48	66
13	6 5	3	2	134	0	134	480	150	18	132
13	6 5	3	3	90	42	48	534	134	94	40
13	6 5	4	3	82	10	72	482	104	14	90
13	6 5	4	4	86	60	26	548	134	104	30
13	6 5	5	4	62	8	54	500	78	22	56
13	6 5	5	5	90	72	18	556	180	168	12

Table 3.4.2

Summary Statistics for Various Tailored Testing
and Conventional Testing Simulations with Hierarchy A
(Totals Based on a Sample of 100 Examinees)

Starting Point	Ability Distribution Parameters N R	Mastery Testing		Tailored Testing				Conventional Testing			
		Length	No. Required For Mastery	Combined Errors	Type I Errors	Type II Errors	No. of Objectives Tested	Combined Errors	Type I Errors	Type II Errors	
13	6	1	1	257	16	241	594	275	59	216	
13	6	2	2	163	38	125	640	208	88	120	
13	6	3	2	234	3	231	555	237	16	221	
13	6	3	3	161	76	85	704	199	126	73	
13	6	4	3	159	11	148	577	159	23	136	
13	6	4	4	154	114	40	776	215	168	47	
13	6	5	4	142	14	128	632	172	35	137	
13	6	5	5	151	115	36	746	220	184	36	
13	6	1	1	178	8	170	484	219	47	172	
13	6	2	2	155	33	122	566	193	82	111	
13	6	3	2	190	3	187	492	184	13	171	
13	6	3	3	132	62	70	594	190	140	50	
13	6	4	3	127	9	118	518	114	16	98	
13	6	4	4	97	59	38	605	189	154	35	
13	6	5	4	116	13	103	540	119	27	92	
13	6	5	5	140	113	27	659	205	173	32	
13	6	1	1	133	15	118	470	151	36	115	
13	6	2	2	98	21	77	500	123	57	66	
13	6	3	2	135	8	127	481	137	10	127	
13	6	3	3	94	45	49	521	125	86	39	
13	6	4	3	80	8	72	485	92	19	73	
13	6	4	4	79	58	21	551	163	133	30	
13	6	5	4	66	11	55	500	78	25	53	
13	6	5	5	110	91	19	570	154	142	12	

Table 3.4.3

Summary Statistics for Various Tailored Testing
and Conventional Testing Simulations with Hierarchy A
(Totals Based on a Sample of 200 Examinees)

Starting Point	Ability Distribution Parameters		Mastery Testing		Tailored Testing				Conventional Testing			
	N	R	Length	No. Required For Mastery	Combined Errors	Type I Errors	Type II Errors	No. of Objectives Tested	Combined Errors	Type I Errors	Type II Errors	
13	6	2	1	1	246	11	235	566	255	46	209	
13	6	2	2	2	177	39	138	639	202	94	108	
13	6	2	3	2	201	2	199	540	192	14	178	
13	6	2	3	3	164	68	96	705	218	136	82	
13	6	2	4	3	156	14	142	598	165	27	138	
13	6	2	4	4	173	115	58	787	204	151	53	
13	6	2	5	4	136	15	121	642	152	38	114	
13	6	2	5	5	169	140	29	791	212	173	39	
13	6	4	1	1	201	15	186	503	211	39	172	
13	6	4	2	2	141	30	111	552	185	78	107	
13	6	4	3	2	169	2	167	494	177	15	162	
13	6	4	3	3	134	66	68	615	187	120	67	
13	6	4	4	3	128	10	118	529	129	20	109	
13	6	4	4	4	127	83	44	627	212	166	46	
13	6	4	5	4	104	10	94	540	121	30	91	
13	6	4	5	5	147	118	29	651	203	174	29	
13	6	5	1	1	127	10	117	465	155	40	115	
13	6	5	2	2	96	26	70	513	136	66	70	
13	6	5	3	2	135	2	133	468	139	11	128	
13	6	5	3	3	81	45	36	529	133	94	39	
13	6	5	3	3								
13	6	5	4	3	86	4	82	483	89	14	75	
13	6	5	4	4	82	59	23	553	142	121	21	
13	6	5	5	4	82	10	72	494	92	26	66	
13	6	5	5	5	88	74	14	573	138	122	16	

Table 3.4.4

Summary Statistics for Various Tailored Testing
and Conventional Testing Simulations with Hierarchy A
(Totals Based on a Sample of 500 Examinees)

Starting Point	Ability Distribution Parameters N R	Mastery Testing		Tailored Testing			Conventional Testing			
		Length	No. Required For Mastery	Combined Errors	Type I Errors	Type II Errors	No. of Objectives Tested	Combined Errors	Type I Errors	Type II Errors
13	6	1	1	263	11	252	577	264	52	212
13	6	2	2	175	40	135	661	202	87	115
13	6	3	2	218	3	215	549	215	19	196
13	6	3	3	156	65	91	714	192	121	71
13	6	4	3	174	8	166	604	178	24	154
13	6	4	4	155	105	50	768	207	154	53
13	6	5	4	137	19	118	647	150	35	115
13	6	5	5	166	131	35	774	214	176	38
13	6	1	1	196	13	183	501	206	41	165
13	6	2	2	143	31	112	549	184	88	96
13	6	3	2	179	3	176	490	183	13	170
13	6	3	3	128	56	72	606	190	125	65
13	6	4	3	140	8	132	522	146	23	123
13	6	4	4	126	84	42	631	192	150	42
13	6	5	4	102	16	86	540	123	35	88
13	6	5	5	144	110	34	660	211	180	31
13	6	1	1	107	12	95	469	149	36	113
13	6	2	2	97	30	67	525	131	63	68
13	6	3	2	114	4	110	469	116	10	106
13	6	3	3	80	45	35	537	128	81	47
13	6	4	3	88	9	79	487	93	13	80
13	6	4	4	91	66	25	564	142	116	26
13	6	5	4	76	13	63	499	85	28	57
13	6	5	5	56	79	17	575	166	148	18

Since the results were essentially identical for the three largest samples, the decision was made to conduct our analyses on a hypothetical sample of 100 examinees. (A complete description of the entries in Tables 3.4.1 to 3.4.4 is provided in section 4.1.)

3.5 Criterion Measures

Two criterion measures were introduced to assess the effectiveness of the tailored testing strategies under different sets of conditions. The first criterion was a count of the number of objectives tested or equivalently, the amount of testing time used. Under conventional testing, 16 objectives were tested in hierarchy A and 18 objectives in hierarchy B. These numbers provided a baseline for interpreting the results with the various tailored testing strategies.

A second criterion was a measure of agreement between an individual's true mastery record and observed mastery record. The number of Type I (false-negative), Type II (false-positive) and combined errors were tabulated for each individual and then totals were compiled by summing across individuals. These totals were compiled for each tailored testing strategy so that comparisons could be made.

3.6 Procedure

A considerable amount of time was spent developing a general CDC 3600 computer program for the purpose of conducting the simulation study. In addition to extensive checks that were made on the accuracy of the computer output, we also applied a two-way chi-square test to establish the appropriateness of the random number generator.

The tailored testing strategies involved the combination of four

starting points, five test lengths, and two observed mastery cutting scores. In addition, we considered three distinct routing procedures. The combinations of the levels of the four factors, starting point, routing procedure, test length, and observed cutting score, were considered with two learning hierarchies and three ability distributions. The two criterion measures for evaluating the tailored testing strategies were the number of objectives tested and the agreement between true and observed mastery records.

The steps involved in our data generation and analysis were as follows:

- Step 1. Pick one of the two hierarchies.
- Step 2. Select one of the four starting points.
- Step 3. Select one of the three ability distributions.
- Step 4. Select one of the five test lengths.
- Step 5. Select one of the two observed mastery cutting scores. (For the 1 and 2 item tests there was only one.)
- Step 6. Select one of the three routing procedures.
- Step 7. For each of 100 individuals proceed through the following steps,
 - a. Select a true mastery score for an individual on objective 13. This was done according to the methodology outlined in Section 3.3 on the distribution of true mastery scores.
 - b. Generate true mastery scores for the individual on the remaining objectives in the hierarchy. Again, this was done according to the methodology outlined in Section 3.3 for the distribution of true mastery scores.

- c. Produce a true mastery record for the individual. That is, for each true mastery score corresponding to an objective in the learning hierarchy, set $\omega = 1$, if $\pi \geq \pi_0$, and set $\omega = 0$, if $\pi < \pi_0$.
- d. Beginning at the appropriate starting point, simulate an individual's performance on n test items measuring the objective. This was done by generating n random numbers from a uniform distribution on the interval $(0, 1)$. The individual's observed mastery score was the proportion of random numbers on the interval $(0, \pi)$ where π was the individual's true mastery score on the objective. Use the observed mastery score to determine the mastery state of the individual. That is, set $\hat{\omega} = 1$, if $\hat{\pi} \geq \hat{\pi}_0$, and set $\hat{\omega} = 0$, if $\hat{\pi} < \hat{\pi}_0$.
- e. Implement the appropriate routing procedure (see step 6). Testing continued until an individual had been assigned to a mastery state for each objective in the learning hierarchy.
- f. Compute the two criterion measures. The first was the number of objectives actually tested. The second was the number of Type I and Type II errors.
- g. Simulate test performance of the individual on all objectives in the learning hierarchy. (This was done by repeating the first part of (d) and making use of the individual's true mastery scores.) Obtain the individual's true and observed mastery record for the purpose of calculating Type

I and Type II errors. These errors provided an indication of the errors that would be obtained in a conventional testing situation.

In summary, for each combination of the levels of the six factors (see Steps 1 to 6) the parts (a) to (g) in Step 7 were carried out to produce some statistics on the effectiveness of a tailored testing strategy under a specified set of conditions.

Chapter IV

Results and Discussion

4.1 Introduction

A summary of the simulation results with the fixed-step-size routing procedure are reported in Tables 4.1.1 and 4.1.2. For convenience we have reported and discussed the results with each of the hierarchies separately. For each combination of the levels of the four factors, starting point, true mastery score distribution, observed mastery cutting score and test length, we have tabulated the number of errors, total errors, Type I errors (false-negative errors) and Type II errors (false-positive errors) and the number of objectives tested across our sample of 100 examinees. In addition we have reported the same information for our simulations of the conventional testing procedure. Totals are reported in the tables for convenience. Also, it should be noted that hierarchy B consisting of 18 objectives is two objectives longer than hierarchy A and this, in part, explains why the number of errors and the number of objectives tested was larger.

The remainder of the chapter is divided into three sections. The first is a comparison of tailored testing and conventional testing results reported in Tables 4.1.1 and 4.1.2. The second is an in-depth analysis of selected features of the simulation results. In the final section we have reported on our comparative analysis of the three routing procedures.

4.2 A Comparison of Tailored Testing and Conventional Testing

The average number of errors of classification for hierarchy A

Table 4.1.1.1

Summary Statistics for Various Tailored Testing
and Conventional Testing Simulations with Hierarchy A
(Totals Based on a Sample of 100 Examinees)

Starting Point	Ability Distribution Parameters N R	Mastery Testing		Tailored Testing				Conventional Testing			
		Length	No. Required For Mastery	Combined Errors	Type I Errors	Type II Errors	No. of Objectives Tested	Combined Errors	Type I Errors	Type II Errors	
5	6	1	1	188	18	170	737	226	45	181	
5	6	2	2	172	75	97	766	200	94	106	
5	6	3	2	176	13	163	741	208	15	193	
5	6	3	3	186	115	71	769	202	128	74	
5	6	4	3	162	13	149	736	177	21	156	
5	6	4	4	178	138	40	763	211	166	45	
5	6	5	4	157	29	128	759	171	37	134	
5	6	5	5	175	156	19	765	212	177	35	
5	6	1	1	190	30	160	737	212	37	175	
5	6	2	2	148	74	74	758	199	95	104	
5	6	3	2	161	13	148	720	162	10	152	
5	6	3	3	143	113	30	753	159	102	57	
5	6	4	3	113	12	101	726	115	16	99	
5	6	4	4	168	148	19	753	204	166	39	
5	6	5	4	110	32	78	731	110	26	84	
5	6	5	5	171	145	26	750	207	170	37	
5	6	1	1	141	25	116	737	137	18	119	
5	6	2	2	102	50	52	736	127	69	58	
5	6	3	2	135	7	128	726	122	10	112	
5	6	3	3	105	91	14	735	126	93	33	
5	6	4	3	86	3	83	722	102	19	83	
5	6	4	4	119	99	20	737	141	122	19	
5	6	5	4	76	20	56	727	86	17	69	
5	6	5	5	175	167	8	734	182	157	25	

Table 4.1.1 (continued)

Summary Statistics for Various Tailored Testing
and Conventional Testing Simulations with Hierarchy A
(Totals Based on a Sample of 100 Examinees)

Starting Point	Ability Distribution Parameters		Mastery Testing		Tailored Testing				Conventional Testing			
	N	R	Length	No. Required For Mastery	Combined Errors	Type I	Type II	No. of Objectives Tested	Combined Errors	Type I	Type II	
						Errors	Errors			Errors	Errors	
10	6	2	1	1	245	12	233	545	274	53	221	
10	6	2	2	2	199	29	170	616	220	74	146	
10	6	2	3	2	181	5	176	537	189	12	177	
10	6	2	3	3	167	74	93	654	212	135	77	
10	6	2	4	3	199	13	186	572	190	29	161	
10	6	2	4	4	140	98	42	675	211	164	47	
10	6	2	5	4	153	14	139	586	168	42	126	
10	6	2	5	5	168	118	50	712	229	180	49	
10	6	4	1	1	164	19	145	532	210	51	159	
10	6	4	2	2	137	40	97	549	181	88	93	
10	6	4	3	2	182	8	174	507	170	7	163	
10	6	4	3	3	137	81	56	596	195	131	64	
10	6	4	4	3	108	11	97	527	136	21	115	
10	6	4	4	4	119	90	29	581	179	138	41	
10	6	4	5	4	113	23	90	549	123	40	83	
10	6	4	5	5	157	134	23	622	219	191	28	
10	6	5	1	1	113	24	89	523	151	37	114	
10	6	5	2	2	81	31	50	520	114	47	67	
10	6	5	3	2	113	5	108	506	118	7	111	
10	6	5	3	3	92	59	33	550	125	85	40	
10	6	5	4	3	95	11	84	510	93	18	75	
10	6	5	4	4	111	95	16	551	130	111	19	
10	6	5	5	4	78	18	60	518	86	21	65	
10	6	5	5	5	128	113	10	565	160	146	14	

Table 4.1.1.1 (continued)
Summary Statistics for Various Tailored Testing
and Conventional Testing Simulations with Hierarchy A
(Totals Based on a Sample of 100 Examinees)

Starting Point	Ability Distribution Parameters N R	Mastery Testing		Tailored Testing			Conventional Testing		
		Length	No. Required For Mastery	Combined Errors	Type I Errors	Type II Errors	Combined Errors	Type I Errors	Type II Errors
13	6	1	1	257	16	241	275	59	215
13	6	2	2	163	38	125	208	83	120
13	6	3	2	234	3	231	237	16	221
13	6	3	3	161	76	85	199	126	73
13	6	4	3	159	11	148	159	23	136
13	6	4	4	154	114	40	215	168	47
13	6	5	4	142	14	128	172	35	137
13	6	5	5	151	115	36	220	184	36
13	6	1	1	178	8	170	219	47	172
13	6	2	2	155	33	122	193	82	111
13	6	3	2	190	3	187	184	13	171
13	6	3	3	132	62	70	190	140	50
13	6	4	3	127	9	118	114	16	93
13	6	4	4	97	59	38	189	154	35
13	6	5	4	116	13	103	119	27	92
13	6	5	5	140	113	27	205	173	32
13	6	1	1	133	15	118	151	36	115
13	6	2	2	93	21	77	123	57	66
13	6	3	2	135	8	127	137	10	127
13	6	3	3	94	45	49	125	86	39
13	6	4	3	80	8	72	92	19	73
13	6	4	4	79	58	21	163	133	30
13	6	5	4	66	11	55	78	25	53
13	6	5	5	110	91	19	154	142	12

Table 4.1.1 (continued)

Summary Statistics for Various Tailored Testing
and Conventional Testing Simulations with Hierarchy A
, (Totals Based on a Sample of 100 Examinees)

Starting Point	Ability Distribution Parameters N R	Mastery Testing		Tailored Testing			Conventional Testing		
		Length	No. Required For Mastery	Combined Errors	Type I Errors	Type II Errors	No. of Objectives Tested	Combined Errors	Type I Errors Type II Errors
16	6 2	1	1	220	14	206	739	238	40 198
16	6 2	2	2	150	35	115	821	187	90 97
16	6 2	3	2	227	3	224	713	212	11 201
16	6 2	3	3	160	72	88	902	187	116 71
16	6 2	4	3	173	9	164	794	166	27 139
16	6 2	4	4	163	103	60	955	208	165 43
16	6 2	5	4	117	15	102	799	113	32 81
16	6 2	5	5	168	121	47	955	196	164 32
16	6 2	5	5						
16	6 4	1	1	201	14	187	630	216	39 177
16	6 4	2	2	141	33	108	708	176	77 99
16	6 4	3	2	177	7	170	627	168	9 159
16	6 4	3	3	128	62	66	762	173	120 53
16	6 4	4	3	142	8	134	664	149	23 126
16	6 4	4	4	116	71	45	818	199	143 56
16	6 4	5	4	119	11	108	679	120	22 98
16	6 4	5	5	149	114	35	854	203	179 24
16	6 5	1	1	144	15	129	575	148	44 104
16	6 5	2	2	104	40	64	660	128	72 56
16	6 5	3	2	112	5	107	586	109	11 98
16	6 5	3	3	79	49	30	671	133	90 43
16	6 5	4	3	101	4	97	597	114	18 96
16	6 5	4	4	105	88	17	744	147	117 30
16	6 5	5	4	60	9	51	621	74	23 51
16	6 5	5	5	109	85	24	727	185	159 26

Table 4.1.2

Summary Statistics for Various Tailored Testing
and Conventional Testing Simulations with Hierarchy B
(Totals Based on a Sample of 100 Examinees)

Starting Point	Ability Distribution Parameters N R	Mastery Testing		Tailored Testing			Conventional Testing		
		Length	No. Required For Mastery	Combined Errors	Type I Errors	Type II Errors	Combined Errors	Type I Errors	Type II Errors
1	6 2	1	1	331	66	265	408	31	377
1	6 2	2	2	254	142	112	279	59	220
1	6 2	3	2	302	45	257	404	11	393
1	6 2	3	3	248	176	72	220	79	141
1	6 2	4	3	252	61	191	293	30	263
1	6 2	4	4	270	232	38	226	131	95
1	6 2	5	4	224	102	122	219	38	181
1	6 2	5	5	300	275	25	212	152	60
1	6 4	1	1	277	123	154	290	46	244
1	6 4	2	2	254	176	78	251	114	137
1	6 4	3	2	194	22	172	210	13	197
1	6 4	3	3	259	239	20	227	156	71
1	6 4	4	3	206	56	150	222	25	197
1	6 4	4	4	291	255	36	248	179	69
1	6 4	5	4	166	100	66	187	57	130
1	6 4	5	5	340	323	17	254	207	47
1	6 5	1	1	184	73	111	181	32	149
1	6 5	2	2	204	169	35	173	99	74
1	6 5	3	2	115	30	85	141	10	131
1	6 5	3	3	253	226	27	187	138	49
1	6 5	4	3	94	40	54	93	31	62
1	6 5	4	4	292	280	12	218	178	40
1	6 5	5	4	120	79	41	102	33	69
1	6 5	5	5	318	309	9	206	181	25

Table 4.1.2 (continued)

Summary Statistics for Various Tailored Testing
and Conventional Testing Simulations with Hierarchy B
(Totals Based on a Sample of 100 Examinees)

Starting Point	Ability Distribution Parameters N R	Mastery Testing		Tailored Testing			Conventional Testing		
		Length	No. Required For Mastery	Combined Errors	Type I Errors	Type II Errors	No. of Objectives Tested	Combined Errors	Type I Errors Type II Errors
13	6 2	1	1	474	19	455	714	390	44 346
13	6 2	2	2	370	62	308	732	277	72 205
13	6 2	3	2	482	16	466	670	383	16 367
13	6 2	3	3	336	87	249	805	237	97 140
13	6 2	4	3	389	32	357	740	299	20 279
13	6 2	4	4	260	117	143	784	204	125 79
13	6 2	5	4	323	42	281	752	234	37 197
13	6 2	5	5	274	185	89	834	213	155 58
13	6 4	1	1	278	22	256	624	250	58 192
13	6 4	2	2	242	51	191	665	277	126 151
13	6 4	3	2	247	10	237	607	219	19 200
13	6 4	3	3	243	92	151	728	239	142 97
13	6 4	4	3	245	10	235	656	231	29 202
13	6 4	4	4	211	108	103	752	214	163 51
13	6 4	5	4	240	26	214	659	201	45 156
13	6 4	5	5	216	167	49	776	242	208 34
13	6 5	1	1	117	15	102	596	148	55 93
13	6 5	2	2	142	44	98	626	150	65 85
13	6 5	3	2	115	3	112	583	117	9 108
13	6 5	3	3	114	67	47	643	157	121 36
13	6 5	4	3	125	15	110	609	120	21 99
13	6 5	4	4	128	92	36	664	187	162 25
13	6 5	5	4	109	20	89	604	105	39 66
13	6 5	5	5	143	102	41	672	209	187 22

Table 4.1.2 (continued)

Summary Statistics for Various Tailored Testing
and Conventional Testing Simulations with Hierarchy B
(Totals Based on a Sample of 100 Examinees)

Starting Point	Ability Distribution Parameters		Mastery Testing		Tailored Testing				Conventional Testing		
	N	R	Length	No. Required For Mastery	Combined Errors	Type I Errors	Type II Errors	No. of Objectives Tested	Combined Errors	Type I Errors	Type II Errors
17	6	2	1	1	461	29	432	963	378	39	339
17	6	2	2	2	345	64	281	1026	263	76	187
17	6	2	3	2	487	21	466	974	385	9	376
17	6	2	3	3	310	78	232	1081	244	113	131
17	6	2	4	3	447	50	397	1037	324	18	306
17	6	2	4	4	277	146	131	1102	183	89	94
17	6	2	5	4	343	47	296	1042	256	45	211
17	6	2	5	5	265	176	89	1126	253	193	60
17	6	4	1	1	285	31	254	924	279	68	211
17	6	4	2	2	290	79	211	988	249	125	124
17	6	4	3	2	264	10	254	911	223	15	208
17	6	4	3	3	240	101	139	1025	253	167	86
17	6	4	4	3	257	8	249	944	214	30	184
17	6	4	4	4	244	138	106	1047	260	182	78
17	6	4	5	4	223	28	195	970	199	46	153
17	6	4	5	5	242	139	103	1042	263	196	67
17	6	5	1	1	155	17	138	881	156	35	121
17	6	5	2	2	172	51	121	911	163	88	75
17	6	5	3	2	113	7	106	893	125	16	109
17	6	5	3	3	136	63	73	956	168	128	40
17	6	5	4	3	107	14	93	898	100	20	80
17	6	5	4	4	157	100	57	972	192	158	34
17	6	5	5	4	112	9	103	905	110	37	73
17	6	5	5	5	165	134	31	979	207	188	19

Table 4.1.2 (continued)

Summary Statistics for Various Tailored Testing
and Conventional Testing Simulations with Hierarchy B
, (Totals Based on a Sample of 100 Examinees)

Starting Point	Ability Distribution Parameters N R	Mastery Testing		Tailored Testing			Conventional Testing		
		Length	No. Required For Mastery	Combined Errors	Type I Errors	Type II Errors	Combined Errors	Type I Errors	Type II Errors
18	6 2	1	1	489	32	457	374	40	334
18	6 2	2	2	382	64	318	291	74	217
18	6 2	3	2	446	13	433	323	16	307
18	6 2	3	3	296	99	197	208	95	113
18	6 2	4	3	377	36	341	276	24	252
18	6 2	4	4	259	134	125	186	117	69
18	6 2	5	4	339	72	267	217	34	183
18	6 2	5	5	256	174	82	234	174	60
18	6 4	1	1	296	24	272	287	69	213
18	6 4	2	2	293	61	232	273	109	164
18	6 4	3	2	264	13	251	219	19	200
18	6 4	3	3	234	88	146	248	141	107
18	6 4	4	3	249	12	237	209	32	177
18	6 4	4	4	228	147	81	258	156	62
18	6 4	5	4	258	32	226	212	49	163
18	6 4	5	5	221	155	66	260	215	45
18	6 5	1	1	144	25	119	159	59	100
18	6 5	2	2	147	49	98	160	92	68
18	6 5	3	2	108	14	94	89	10	79
18	6 5	3	3	124	65	59	160	103	52
18	6 5	4	3	125	7	118	106	16	90
18	6 5	4	4	134	88	46	193	172	21
18	6 5	5	4	103	18	85	102	38	64
18	6 5	5	5	151	140	11	216	197	19

across all of the tailored testing strategies and individuals was 1.43. For the conventional testing simulations the average was 1.68. In addition, the average number of objectives tested for hierarchy A across all of the tailored testing strategies and individuals was 6.54. With the conventional testing simulations, individuals were tested on all 16 objectives in the hierarchy. Thus, not only did the tailored testing strategies result in slightly fewer errors of classification, on the average, than conventional testing, but only 40.8% as much testing time was required. To say it in another way, on the average, the tailored testing strategies resulted in a 59.2% reduction in testing time. (At first glance it may seem paradoxical that the errors of classification were greater with conventional testing even with the additional testing time. This can be explained when one considers the fact that more than twice as many test items were encountered during conventional testing thereby increasing the probability of making additional errors of classification.)

The analysis was repeated for hierarchy B and similar results were obtained. The average number of errors of classification across all of the tailored testing strategies and individuals was 2.49 and the average number of objectives tested was 8.43. In contrast, the average number of errors of classification with the conventional testing procedures was 2.26 with a total of 18 objectives tested. With hierarchy B, for the tailored testing strategies there was a 53.2% reduction in testing time with the average number of errors of classification slightly higher.

4.3 Investigation of the Importance of the Factors of Starting Point and Ability Distribution on the Number of Errors of Classification

The data for the analysis were reported in Table 4.3.1 for hierarchy A and Table 4.3.2 for hierarchy B. By way of explanation for Tables 4.3.1 and 4.3.2 and several tables later in the chapter we note that the average number of errors with each starting point and ability distribution was obtained by summing over simulations of different test lengths and cutting scores. Numbers reported in the "Range of Classification Errors" column are the minimum and maximum number of classification errors for each ability distribution across all combinations of the remaining factors.

The results for hierarchy A indicated that the average values of the errors of classification varied only slightly as a function of starting point. The very slight trend was that with a high negatively skewed distribution of true mastery scores, the number of errors of classification were smaller if testing was initiated from the top rather than the bottom of the learning hierarchy.

This trend was much more pronounced with hierarchy B. As expected, with a positively-skewed distribution of true mastery scores it was more efficient to initiate testing at the bottom of the hierarchy. On the average about one less error of classification per individual was made. With a high negatively-skewed distribution of true mastery scores, on the average, about 1/2 fewer errors of classification per individual were made by initiating testing at the top rather than the bottom of the hierarchy.

An additional finding revealed by the tables was that the average number of errors of classification tended to be largest for the positively

Table 4.3.1

Summary of the Average Number of Errors of Classification for
Combinations of Levels of Ability Distribution and
Starting Point Factors with Hierarchy A
(Totals Based on a Sample of 100 Examinees)

Ability Distribution R/N	Range of Classification Errors	Average Number of Errors with Starting Point			
		5	10	13	16
2/6	117-257	174	182	178	173
4/6	97-201	151	140	142	147
5/6	60-175	117	101	99	102

Table 4.3.2

Summary of the Average Number of Errors of Classification for
Combinations of Levels of Ability Distribution and
Starting Point Factors with Hierarchy B
(Totals Based on a Sample of 100 Examinees)

Ability Distribution R/N	Range of Classification Errors	Average Number of Errors with Starting Point			
		1	13	17	18
2/6	224-489	274	377	367	356
4/6	166-340	249	241	256	256
5/6	94-318	197	124	139	130

skewed distribution of true mastery scores and lowest for the high negatively skewed distributions.

4.4 Investigation of the Importance of the Factors of Starting Point and Ability Distribution on the Number of Objectives Tested

Tables 4.4.1 and 4.4.2 summarized the data for this analysis. The tables were constructed in the same way as Tables 4.3.1 and 4.3.2 with the average number of objectives tested substituted for the average number of classification errors.

The results of this analysis were fairly clear although there were two exceptions. In summary it could be said that the number of objectives tested was fewer when middle level objectives were chosen as the starting point. The reduction was about two objectives per individual for hierarchy A and in the range of two to three objectives per individual for hierarchy B. (With hierarchy A, middle level objectives were numbers 10 and 13 and with hierarchy B, objective number 13 was the only middle level objective.)

4.5 Investigation of the Importance of the Factors of Test Length, Cutting Score, and Ability Distribution on the Number of Errors of Classification

The data for this analysis were reported for hierarchy A in Table 4.5.1 and for hierarchy B in Table 4.5.2. The results in Table 4.5.1 for hierarchy A revealed that the average number of errors of classification was lowest for the five item test, with a mastery cutting score of 0.80, regardless of the ability distribution. Similar results were obtained with hierarchy B. (One exception occurred with the positively-skewed distribution of true mastery scores.)

Table 4.4.1

Summary of the Average Number of Objectives Tested for
Combinations of Levels of Ability Distribution and
Starting Point Factors with Hierarchy A
(Totals Based on a Sample of 100 Examinees)

Ability Distribution R/N	Range of Objectives Tested	Ave. No. of Objs. Tested with Starting Point			
		5	10	13	16
2/6	537-955	755	611	653	835
4/6	484-854	741	558	558	780
5/6	470-744	732	530	510	649

Table 4.4.2

Summary of the Average Number of Objectives Tested for
Combinations of Levels of Ability Distribution and
Starting Point Factors with Hierarchy B
(Totals Based on a Sample of 100 Examinees)

Ability Distribution R/N	Range of Objectives Tested	Ave. No. of Objs. Tested with Starting Point			
		1	13	17	18
2/6	531-1126	689	755	1043	976
4/6	607-1047	820	684	983	891
5/6	583-979	885	625	924	834

Table 4.5.1

Summary of the Average Number of Errors of Classification for
Combinations of Levels of Ability Distribution, Cutting Score and
Test Length Factors with Hierarchy A
(Totals Based on a Sample of 100 Examinees)

Ability Distribution R/N	Average Number of Errors							
	Number of Correct Answers Required for Mastery				Total Number of Test Items			
	1/1	2/2	2/3	3/3	3/4	4/4	4/5	5/5
2/6	227	172	204	168	173	159	142	165
4/6	183	145	177	135	122	125	115	155
5/6	133	94	124	142	91	103	70	130

Table 4.5.2

Summary of the Average Number of Errors of Classification for
Combinations of Levels of Ability Distribution, Cutting Score and
Test Length Factors with Hierarchy B
(Totals Based on a Sample of 100 Examinees)

Ability Distribution R/N	Average Number of Errors							
	Number of Correct Answers Required for Mastery				Total Number of Test Items			
	1/1	2/2	2/3	3/3	3/4	4/4	4/5	5/5
2/6	439	338	429	298	366	267	307	274
4/6	284	270	242	244	239	243	222	254
5/6	150	166	113	157	113	178	111	194

As one moved from a test length of one item and a mastery cutting score of 1.0 to a test length of five items and a mastery cutting score of 0.80 the average number of errors of classification was reduced by about 40%. Also, as was the case with other analyses it was revealed that the average number of errors of classification was usually lowest for the high negatively-skewed distribution of true mastery scores and highest for the positively-skewed distribution of true mastery scores.

4.6 Investigation of the Importance of the Factors of Test Length, Cutting Score, and Ability Distribution on the Number of Objectives Tested

The data for this analysis were reported for hierarchy A in Table 4.6.1 and for hierarchy B in Table 4.6.2. The results for hierarchy A indicated that the fewest number of objectives tested occurred, on the average, at a test length of three items with an observed mastery cutting score of 0.67. The maximum number of objectives tested occurred, on the average, with tests of five items and an observed mastery cutting score of 1.0. With hierarchy B there was surprisingly little variation in the numbers of objectives tested as a function of ability distribution, cutting score, and test length. But what differences occurred were in the same locations and in the same direction as with hierarchy A.

4.7 Investigation of the Importance of the Factors of Test Length, Cutting Score, and Ability Distribution on the Percentage of Type I Errors of Classification

The data for our analysis of hierarchy A were reported in Table 4.7.1 and for hierarchy B in Table 4.7.2. Our results indicated two significant trends in the data which were clear from both tables. The first was that as one

Table 4.6.1

Summary of the Average Number of Objectives Tested for
Combinations of Levels of Ability Distribution, Cutting Score and
Test Length Factors with Hierarchy A
(Totals Based on a Sample of 100 Examinees)

Ability Distribution R/N	Average Number of Objectives Tested							
	Number of Correct Answers/Total Number of Required for Mastery / Test Items							
	1/1	2/2	2/3	3/3	3/4	4/4	4/5	5/5
2/6	654	711	637	758	669	792	694	795
4/6	596	645	587	676	609	689	624	721
5/6	576	603	575	619	578	646	592	649

Table 4.6.2

Summary of the Average Number of Objectives Tested for
Combinations of Levels of Ability Distribution, Cutting Score and
Test Length Factors with Hierarchy B
(Totals Based on a Sample of 100 Examinees)

Ability Distribution R/N	Average Number of Objectives Tested							
	Number of Correct Answers/Total Number of Required for Mastery / Test Items							
	1/1	2/2	2/3	3/3	3/4	4/4	4/5	5/5
2/6	841	858	828	885	880	864	877	896
4/6	809	847	818	866	834	879	831	876
5/6	804	806	802	826	805	833	807	842

Table 4.7.1

Summary of the Average Percent of Type I Errors of Classification for Combinations of Levels of Ability Distribution, Cutting Score and Test Length Factors with Hierarchy A

Ability Distribution R/N	Average Percent of Type I Errors of Classification							
	Number of Correct Answers/ Required for Mastery					Total Number of Test Items		
	1/1	2/2	2/3	3/3	3/4	4/4	4/5	5/5
2/6	6.6	16.7	2.9	49.9	6.6	71.4	12.7	77.1
4/6	9.7	30.9	3.3	58.7	6.3	73.7	12.9	82.1
5/6	14.9	36.9	4.4	65.8	8.2	82.2	17.3	88.4

Table 4.7.2

Summary of the Average Percent of Type I Errors of Classification for Combinations of Levels of Ability Distribution, Cutting Score and Test Length Factors with Hierarchy B

Ability Distribution R/N	Average Percent of Type I Errors of Classification							
	Number of Correct Answers/ Required for Mastery					Total Number of Test Items		
	1/1	2/2	2/3	3/3	3/4	4/4	4/5	5/5
2/6	8.3	24.5	5.5	37.0	12.2	59.0	21.4	74.0
4/6	17.6	33.9	6.6	53.2	9.3	66.5	19.3	76.9
5/6	21.7	47.1	5.7	67.1	9.0	78.7	20.9	88.2

moved from a positively skewed distribution to a high negatively skewed distribution of true mastery scores the percentage of Type I errors of the total errors of classification increased. Second, it was quite clear that as the test length was increased, for a fixed mastery cutting score of 1.0, that the percentage of Type I errors of the total number of errors of classification increased irrespective of the ability distribution. The percentage varied from about 10% for the one-item tests to over 80% for the five-item tests. Finally it was quite clear from the data that if the observed mastery cutting score was set below the true mastery cutting score (i.e., when the mastery cutting was set at .67 and .75) that the percentage of Type I errors of the total number of errors was low. The reverse was true with an observed mastery cutting scores above the true mastery cutting score. The larger percentage of the number of errors of classification were of the Type II variety.

Thus these data provide some very useful information on the effect of the observed mastery cutting score on the type and number of errors of classification.

4.8 A Comparison of the Effectiveness of Three Routing Procedures

The data for this analysis were reported for hierarchy A in Tables 4.8.1 and 4.8.3 and for hierarchy B in Tables 4.8.2 and 4.8.4. Three variations in routing strategy were investigated. Two of the methods, referred to as method I and II, involved variable-step-size procedures while the third method involved the fixed-step-size procedure used in the other parts of our analysis. Complete details of the three routing

Table 4.8.1

Summary of the Average Number of Errors of Classification
for Three Levels of the Ability Distribution Factor and
Three Routing Strategies with Hierarchy A
(Totals Based on a Sample of 100 Examinees)

Ability Distribution R/N	Average Number of Errors of Classification		
	Method I	Method II	Method III
2/6	200	177	178
4/6	166	152	142
5/6	104	101	99

Table 4.8.2

Summary of the Average Number of Errors of Classification
for Three Levels of the Ability Distribution Factor and
Three Routing Strategies with Hierarchy B
(Totals Based on a Sample of 100 Examinees)

Ability Distribution R/N	Average Number of Errors of Classification		
	Method I	Method II	Method III
2/6	301	324	377
4/6	268	251	241
5/6	179	152	124

Table 4.8.3

Summary of the Average Number of Objectives Tested
for Three Levels of the Ability Distribution
Factor and Three Routing Strategies with Hierarchy A
(Totals Based on a Sample of 100 Examinees)

Ability Distribution R/N	Average Number of Objectives Tested		
	Method I	Method II	Method III
2/6	773	711	653
4/6	698	617	558
5/6	625	583	510

Table 4.8.4

Summary of the Average Number of Objectives Tested
for Three Levels of the Ability Distribution
Factor and Three Routing Strategies with Hierarchy B
(Totals Based on a Sample of 100 Examinees)

Ability Distribution R/N	Average Number of Objectives Tested		
	Method I	Method II	Method III
2/6	740	752	755
4/6	831	753	684
5/6	872	750	625

procedures were described in Section 3.3. The analyses were conducted with a test length of five items, an observed mastery cutting score of 0.8 and objective 13 as the starting point.

The results were rather different for the two hierarchies. With hierarchy A, the results indicated that regardless of ability distribution method III was better than method II and method II was better than method I. In terms of the average number of errors of classification however the differences were slight. The differences were somewhat larger among the three routing methods for the numbers of objectives tested.

For hierarchy B, method I and II produced the fewest average number of errors of classification and objectives tested for the positively skewed distribution of true mastery scores across the objectives in the learning hierarchy for the sample of individuals. On the other hand, substantially better results were obtained from the high negatively skewed distribution of true mastery scores with method III. In all cases the difference between method I and III were the most pronounced.

Chapter V

Conclusions

5.1 Concluding Remarks

The development and implementation of objective-based programs in education has been proceeding at a rapid rate. One of the major criticisms of these programs is that they require an extensive amount of testing time. Of the many suggestions that have been made for reducing the testing time, one of the most appealing is to implement some version of tailored testing. Unfortunately there has been only one study (Ferguson, 1969) reported in the literature that would provide any guidelines for how this might be done. Hence there is considerable need for further studies on the various facets of tailored testing before an implementation phase should take place. It should be noted that the results of Ferguson's work apply only to computer-administered tests whereas at the present time, such computer facilities are not typically used for this purpose in objective-based programs.

The purpose of this simulation study was to assess the effectiveness of various tailored testing strategies for different testing situations. The four factors of a tailored testing strategy under study were test length, mastery cutting score, starting point and the routing procedure. The various levels of the four factors were combined to generate a multitude of tailored testing strategies that were studied for their effectiveness at reducing the testing time and errors of classification with two learning hierarchies and three different distributions

of true mastery scores for individuals across the hierarchies. Primarily because of the complexity of the problem and our desire to produce a large quantity of data on tailored testing strategies, a computer simulation approach to the problem was decided upon. Since our primary goal was to produce results of tailored testing strategies that would provide guidelines for future empirical studies and implementation in objective-based programs operating without the aid of computer terminals, only simple tailored testing strategies were considered. Primarily this meant that decision rules for monitoring an individual through a hierarchy were simple enough so that they could be implemented in a classroom testing situation.

Our results indicated that it was possible to obtain an overall reduction of more than 50% in testing time which could be achieved by utilizing a tailored testing procedure as an alternative to a conventional testing procedure. With Gagné's hierarchy (hierarchy A) the tailored testing strategies resulted, on the average, in an overall reduction of testing time of 59.2%. With Ferguson's hierarchy (hierarchy B) the reduction was by 53.2%. It is likely that tailored testing strategies with hierarchy B were not quite as effective as with hierarchy A because of the fact that hierarchy B had two terminal objectives whereas hierarchy A had only one. The difference in results serve to highlight the importance of the particular form of the structure of objective on the effectiveness of tailored testing strategies. Nevertheless it was dramatically clear that there was a considerable saving in testing time to be gained by resorting to a tailored testing

strategy, regardless of the hierarchy. In the future perhaps more sophisticated strategies might be produced to more effectively handle hierarchies with more than one terminal objective.

We note that with certain tailored testing strategies the reduction in testing time was as much as 70%. This result was achieved with both hierarchies. Our results on the saving of testing time which varied from 50-70% compared favorably with the empirical results of Ferguson (1969) where he reported a saving of testing time of 60% over conventional testing. The similarity of the results would seem to add some validity to the appropriateness of our simulation procedures.

While the reductions in testing time derived from the tailored testing strategies were impressive they would have meant little if the total numbers of errors of classification were substantially larger. Fortunately this was not the case. In fact with hierarchy A the tailored testing strategies resulted in a slightly lower number of errors of classification than with conventional testing. The reverse was true with hierarchy B but again, the differences were slight. This finding along with the information on the comparisons of testing time for conventional and tailored testing provide strong support for what many researchers have been speculating on. That is, there is much to be gained in terms of testing time efficiency without any significant loss in the accuracy of decision-making by resorting to a tailored testing strategy in the context of objective-based programs.

Looking next at the various tailored testing strategies we note several general findings. First, if the distribution of true mastery

scores across the objectives in the learning hierarchy was positively skewed, it was far more effective, in terms of reducing the total number of errors of classification, to initiate testing at the bottom part of the hierarchy. With a high negatively skewed distribution of true mastery scores, the total number of errors of classification was reduced by initiating testing at the top rather than the bottom of the learning hierarchy. However, it should be noted that what differences were obtained tended to be small.

Second, on the matter of testing time, it seemed that testing time was lowest when testing was initiated in the middle rather than at the extremes of the learning hierarchy. About two less objectives, on the average, were tested if testing was initiated in the middle of the learning hierarchy.

Third, on the average, there was a 40% reduction in the number of errors of classification with a five-item test and an observed mastery cutting score of 0.80 as compared to a one-item test with an observed mastery cutting of 1.00. As expected, the longer the test, the fewer the average number of errors of classification.

Fourth, a study of the types of errors (false-positive and false-negative) revealed that this was controlled by the position of the observed cutting score. When the observed cutting score was set above the true mastery cutting score, the majority of the errors were of the false-negative type. If it was set lower, the highest percentage of the total errors were of the false-positive type.

Finally we note that while our analysis of routing methods was not extensive, it did indicate the importance of routing methods on the

overall effectiveness of a tailored testing strategy and with two different learning hierarchies. Clearly though this is one of the important areas for further study.

5.2 Suggestions for Further Research

Among the problems that remain to be resolved in the area of tailored testing research, two seem particularly important. The first involves an extension of the work described in this study. Of most importance we see a need for further study of routing methods and stopping rules. In addition it would likely be useful to consider test models in the simulation of test data that incorporated a guessing factor since it is well-known that guessing does play some part in individual test performance.

A second line of research would involve some empirical research on tailored testing in the schools. The design of such a study would involve developing a programmed instruction booklet which would include test items designed to measure specific objectives in a learning hierarchy, a self-scoring device and routing directions. Among the factors that could be investigated in an empirical study are test length, mastery cutting score, and routing method. In addition it would be interesting to study the merits, in terms of overall testing efficiency, of having individuals generate their own starting points for testing in the learning hierarchy.

References

- Airasian, P.W. Formative evaluation instruments. The Irish Journal of Education, 1968, 2, 1927-135.
- Airasian, P.W. & Bart, W.M. Tree theory--A theory generative measurement model. Paper presented at the annual meeting of A.E.R.A., New York, 1971.
- Airasian, P.W. The use of hierarchies in curriculum analysis and instructional planning. California Journal of Educational Research, 1971, 22, 34-41.
- Atkinson, R.C. Computer-based instruction in initial reading. In Proceedings of the 1967 Invitational Conference on Testing Problems. Princeton, N.J.: Educational Testing Service, 1968.
- Atkinson, R.C., & Wilson, H.A. (Eds.) Computer-assisted instruction: A book of readings. New York: Academic Press, 1969.
- Bayroff, A.G., Thomas, J.J. & Anderson, A.A. Construction of an experimental sequential item test. Research Memorandum 60-1, Personnel Research Branch, Department of the Army, January 1960.
- Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord and M.R. Novick, Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.
- Block, J.H. (Ed.) Mastery learning: Theory and practice. New York: Holt, Rinehart, and Winston, 1971.
- Bloom, B.S. Learning for mastery. Evaluation Comment, 1968, 1(2).
- Carroll, J.B. A model of school learning. Teachers College Record, 1963, 64, 723-733.
- Carroll, J.B. Problems of measurement related to the concept of learning for mastery. Educational Horizons, 1970, 48, 71-80.
- Cleary, T.A., Linn, R.L., & Rock, D.A. Exploratory study of programmed tests. Educational and Psychological Measurement, 1968, 28, 345-360.
- Cleary, T.A., Linn, R.L. & Rock, D.A. Reproduction of total test score through the use of sequential programmed tests. Journal of Educational Measurement, 1968, 5, 183-187.

- Cowden, D.J. An application of sequential sampling to testing students. Journal of the American Statistical Association, 1946, 41, 547-556.
- Cronbach, L.J. New light on test strategy from decision theory. In A. Anastasi (Ed.), Testing problems in perspective. Washington, D.C.: American Council on Education, 1966.
- Cronbach, L.J. & Gleser, G.C. Psychological tests and personnel decisions. (2nd ed.) Urbana: University of Illinois Press, 1965.
- DeVault, M.V., Kriewall, T.E., Buchanan, A.E., & Quilling, M.R. Teacher's manual: Computer management for individualized instruction in mathematics and reading. Madison, Wisconsin: University of Wisconsin, Research and Development Center for Cognitive Learning, 1969.
- Ferguson, R.L. The development of a computer-assisted branched test for a program of individually prescribed instruction. Unpublished doctoral dissertation, University of Pittsburgh, 1969.
- Ferguson, R.L. Computer assistance for individualizing instruction. Computers and Automation, 1970, 19, 27-29.
- Ferguson, R.L. Computer assistance for individualizing measurement. Technical Report, Learning Research and Development Center, University of Pittsburgh, March 1971.
- Flanagan, J.C. Functional education for the seventies. Phi Delta Kappan, 1967, 49, 27-32.
- Flanagan, J.C. Program for learning in accordance with needs. Psychology in the Schools, 1969, 6, 133-136.
- Flaugher, R.L. & Rock, D.A. A fixed length optimal test battery for colleges characterized by diverse aptitude levels. American Educational Research Journal, 1968, 5, 659-674.
- Gagné, R.M. The condition of learning. New York: Holt, Rinehart, & Winston, 1965.
- Glaser, R. Adapting the elementary school curriculum to individual performance. In Proceedings of the 1967 Invitational Conference on Testing Problems. Princeton, N.J.: Educational Testing Service, 1968.
- Glaser, R. Evaluation of instruction and changing educational models. In Wittrock, M.C. and Wiley, D.E. (Eds.), The evaluation of instruction. New York: Holt, Rinehart, & Winston, 1970.

- Glaser, R., & Nitko, A.J. Measurement in learning and instruction. In R.L. Thorndike (Ed.), Educational measurement. Washington: American Council on Education, 1971, 625-670.
- Green, B.F., Jr. Comments on tailored testing. In W.H. Holtzman, (Ed.), Computer-assisted instruction, testing, and guidance. New York: Harper & Row, 1970. Pp. 184-197.
- Hambleton, R.K. A review of testing and decision-making procedures for selected individualized instructional programs. ACT Technical Bulletin No. 15. Iowa City, Iowa: American College Testing Program, 1973.
- Hambleton, R.K., & Novick, M.R. Toward an integration of theory and method for criterion-referenced tests. Journal of Educational Measurement. 1973, 10, 159-170.
- Harmon, H.H., Helm, C.E. & Loye, D.E. Computer-assisted teaching. In Proceedings of a Conference on CAT. Princeton, N.J.: Educational Testing Service, 1968.
- Holtzman, W. (Ed.) Individually tailored testing: Discussion. Computer-assisted instruction, testing, and guidance. New York: Harper & Row, 1970. Pp. 198-200.
- Lindvall, C.M. & Cox, R.C. The IPI evaluation program. AERA Monograph Series on Curriculum Evaluation 5. Chicago: Rand McNally and Company, 1970.
- Linn, R.L., Rock, D.A., & Cleary, T.A. The development and evaluation of several programmed testing methods. Educational and Psychological Measurement, 1969, 29, 129-146.
- Linn, R.L., Rock, D.A., & Cleary, T.A. Sequential testing for dichotomous decisions. Research Bulletin 70-31. Princeton, N.J.: Educational Testing Service, 1970.
- Lord, F.M. Review of H. Solomon's Studies in item analysis and prediction. Psychometrika, 1962, 27, 207-213.
- Lord, F.M. Some test theory for tailored testing. In W.H. Holtzman (Ed.), Computer-assisted instruction, testing, and guidance. New York: Harper & Row, 1970. Pp. 139-183.
- Lord, F.M. Robbins-Monro procedures for tailored testing. Educational and Psychological Measurement, 1971, 31, 3-31. (a)
- Lord, F.M. The self-scoring flexilevel test. Journal of Educational Measurement, 1971, 8, 147-151. (b)

- Lord, F.M. Tailored testing, an application of stochastic approximation. Journal of the American Statistical Association, 1971, 66, 707-711. (c)
- Lord, F.M. A theoretical study of the measurement effectiveness of flexilevel tests. Educational and Psychological Measurement, 1971, 31, 805-813. (d)
- Lord, F.M. A theoretical study of two-stage testing. Psychometrika, 1971, 36, 227-242. (e)
- Lord, F.M. & Novick, M.R. Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968. Chapter 16.
- Moonan, W.J. Some empirical aspects of the sequential analysis technique as applied to an achievement examination. Journal of Experimental Education, 1950, 18, 195-207.
- Owen, R.J. A Bayesian approach to tailored testing. Research Bulletin 69-92. Princeton, N.J.: Educational Testing Service, 1969.
- Popham, W.J., & Husek, T.R. Implications of criterion-referenced measurement. Journal of Educational Measurement, 1969, 6, 1-9.
- Rovinelli, R., & Hambleton, R.K. Some procedures for the validation of Criterion-referenced test items. (Final Report) Albany, N.Y.: Bureau of School and Cultural Research, New York State Education Department, 1973.
- Smith, R.E. Examination by computer. Behavioral Science, 1963, 8 76-79.
- Stocking, M. Short tailored tests. Research Bulletin 69-63 and ONR Technical Report, Contract N00014-69-0017. Princeton, N.J.: Educational Testing Service, 1969.
- Suppes, P. The uses of computers in education. Scientific American, 1966, 215, 206-221.
- Urry, V.W. Individualized testing by Bayesian estimation. Bureau of Testing, University of Washington, Seattle, 1971.
- Wald, A. Sequential analysis. New York: Wiley, 1947.
- Washburne, C.W. Educational measurements as a key to individualizing instruction and promotions. Journal of Educational Research, 1922, 5, 195-206.
- Waters, C.W. and Bayroff, A.G. A comparison of computer-simulated conventional and branching tests. Educational and Psychological Measurement, 1971, 31, 125-136.

Weiss, D.J. & Bentz, N.E. Ability measurement: Conventional or adaptive? Research Report 73-1, Psychometric Methods Program, Department of Psychology, University of Minnesota, February 1973.

Wilhelms, F.T. The curriculum and individual differences. In National Society for the Study of Education, Individualizing Instruction, Sixty-First Yearbook, Part I, 1962, 72-82.

Wood, R. The efficacy of tailored testing. Educational Research, 1969, 1, 219-222.

Wood, R. Computerized adaptive sequential testing. Unpublished doctoral dissertation, University of Chicago, 1971.