

1-1-1999

A comparison of computerized adaptive testing and multi-stage testing.

Liane N. Patsula

University of Massachusetts Amherst

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

Recommended Citation

Patsula, Liane N., "A comparison of computerized adaptive testing and multi-stage testing." (1999). *Doctoral Dissertations 1896 - February 2014*. 3282.

https://scholarworks.umass.edu/dissertations_1/3282

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.



A COMPARISON OF COMPUTERIZED ADAPTIVE TESTING AND
MULTI-STAGE TESTING

A Dissertation Presented

By

LIANE N. PATSULA

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

DOCTOR OF PHILOSOPHY

September 1999

Department of Psychology

© Copyright Liane N. Patsula 1999

All Rights Reserved

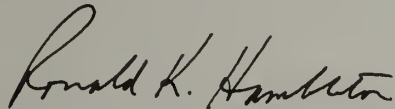
A COMPARISON OF COMPUTERIZED ADAPTIVE TESTING AND
MULTI-STAGE TESTING

A Dissertation Presented

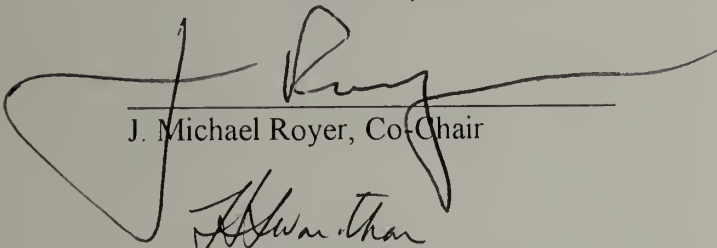
By

LIANE N. PATSULA

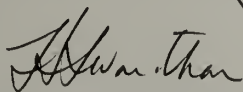
Approved as to style and content by:



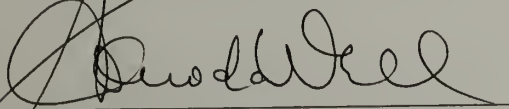
Ronald K. Hambleton, Co-Chair



J. Michael Royer, Co-Chair



Hariharan Swaminathan, Member



Arnold Well, Member



Melinda Novak, Chair
Department of Psychology

ACKNOWLEDGEMENT

While working on this dissertation, I have benefited from the kindness and support of many people. My sincerest gratitude goes to Professor Ronald Hambleton. Not only has he generously shared his insight and enthusiasm, but also his concern and interest for my well-being outside of academics. His quiet concern for where I was living, if I was going to take holidays, and for my finances was very comforting and appreciated. His personal and professional support have been invaluable.

Special thanks is also extended to my committee – Professors Hariharan Swaminathan, J. Michael Royer, and Arnold Well for their thought-provoking questions, ideas, and constructive criticism. I also am greatly indebted to Frederic Robin and Dr. Ric’ Luecht for their software and programming assistance, which contributed immensely to the completion of this dissertation. Thank you. As well, I would like to thank the Research and Evaluation Methods Program for its generous financial support throughout my years at UMass.

Among my graduate student colleagues, I must express particular gratitude to Sharon Cadman Slater for conversation, psychometrically and otherwise, and for the comfort of friendship. All of the REMP students have been very supportive and encouraging. Working alongside them has made the graduate school experience much more enjoyable.

Furthermore, I would especially like to recognize and thank Sean for his never-ending patience with me in spite of my constant preoccupation with the dissertation and

my long hours of seclusion with the computer when I would block out the world. As well, his faith that I could and can do it was great motivation. Having more time together for us and our family is the best reward I can think of for seeing this dissertation to the end.

Above all, I am grateful to my mother and father for what they have given to me and for what they have taught me and continue to teach me. Their unfailing love, support, and understanding have always been a source of inspiration and encouragement. I hope to be able to offer the same to my children. Mom and Dad, I dedicate this dissertation to you in thanksgiving for all that you have given to and done for me.

ABSTRACT

COMPARISON OF COMPUTERIZED ADAPTIVE TESTING
AND MULTI-STAGE TESTING

SEPTEMBER 1999

LIANE PATSULA, B.COM., MCGILL UNIVERSITY, CANADA

M.A., UNIVERSITY OF OTTAWA, CANADA

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professors Ronald K. Hambleton and J. Michael Royer

There is considerable evidence to show that computerized-adaptive testing (CAT) and multi-stage testing (MST) are viable frameworks for testing. With many testing organizations looking to move towards CAT or MST, it is important to know what framework is superior in different situations and at what cost in terms of measurement. What was needed is a comparison of the different testing procedures under various realistic testing conditions. This dissertation addressed the important problem of the increase or decrease in accuracy of ability estimation in using MST rather than CAT.

The purpose of this study was to compare the accuracy of ability estimates produced by MST and CAT while keeping some variables fixed and varying others. A simulation study was conducted to investigate the effects of several factors on the accuracy of ability estimation using different CAT and MST designs. The factors that were manipulated are the number of stages, the number of subtests per stage, and the number of items per subtest. Kept constant were test length, distribution of subtest

information, method of determining cut-points on subtests, amount of overlap between subtests, and method of scoring total test. The primary question of interest was, given a fixed test length, how many stages and many subtests per stage should there be to maximize measurement precision? Furthermore, how many items should there be in each subtest? Should there be more in the routing test or should there be more in the higher stage tests?

Results showed that, in general, increasing the number of stages from two to three decreased the amount of errors in ability estimation. Increasing the number of subtests from three to five increased the accuracy of ability estimates as well as the efficiency of the MST designs relative to the P&P and CAT designs at most ability levels (-.75 to 2.25). Finally, at most ability levels (-.75 to 2.25), varying the number of items per stage had little effect on either the resulting accuracy of ability estimates or the relative efficiency of the MST designs to the P&P and CAT designs.

TABLE OF CONTENTS

	Page
ACKNOWLEDGEMENT	iv
ABSTRACT	vi
LIST OF TABLES.....	xi
LIST OF FIGURES	xii
CHAPTER	
1. INTRODUCTION.....	1
1.1 Background	1
1.2 Multi-Stage Testing.....	7
1.3 Statement of the Problem	9
1.4 Purpose of the Study.....	13
1.5 Significance of Problem	14
2. REVIEW OF LITERATURE.....	15
2.1 The Importance of Item Response Theory in Computerized Adaptive Testing	15
2.2 Computerized Adaptive Testing.....	17
2.2.1 Examples of CAT	18
2.2.2 Advantages of CAT	19
2.2.3 Disadvantages to CAT	22
2.3 Multi-Stage Testing.....	27
2.3.1 Examples of Multi-Stage Testing	28
2.3.2 Advantages of MST	30
2.3.3 Disadvantages of MST.....	34
2.4 Comparison of CAT and MST	35
2.4.1 Summary.....	39
2.5 MST Design Factors.....	39
2.6 Purpose of Study	46

3. METHODOLOGY	48
3.1 Test Conditions	48
3.1.1 P&P Condition.....	50
3.1.2 CAT Condition	50
3.1.3 MST Conditions.....	52
3.1.3.1 Number of Stages	53
3.1.3.2 Number of Subtests per Stage	53
3.1.3.3 Number of Items per Subtest.....	54
3.1.3.4 Summary	55
3.2 Computer Programs.....	56
3.3 Procedure.....	57
3.3.1 Step 1 – Partition Item Pool.....	57
3.3.2. Step 2 – Construct Tests	58
3.3.3. Step 3 – Simulate Examinees	62
3.3.4. Step 4 – Simulate Item Responses.....	63
3.4 Data Analysis.....	63
3.4.1 Ability Estimation.....	72
3.4.2 Item Exposure	74
4. RESULTS.....	75
4.1 Ability Estimation	75
4.1.1 Accuracy	75
4.1.2 Bias	83
4.1.3 Relative Efficiency.....	85
4.1.4 Summary	91
4.2 Item Exposure.....	92
4.2.1 Number of Items Exposed	92
4.2.2 Conditional Exposure Rates	96
4.2.3 Summary	98
4.3 Summary	98

5. CONCLUSION	100
5.1 Conclusion.....	100
5.2 Future Research.....	103
REFERENCES.....	105

LIST OF TABLES

Table	Page
3.1 Content Specifications for all Test Designs	49
3.2 Example of Item Selection	52
3.3 Proportion of Items Per Subtest Per Stage	55
3.4 Number of Items Per Subtest Per Stage	55
3.5 Summary of MST Designs	56
3.6 Descriptive Statistics of Item Parameters in the Total Pool and Each Subpool.....	58
3.7 CASTISEL Assembly Results for P&P and MST Designs.....	64
4.1 Efficiency of MST Relative to CAT	89
4.2 Efficiency of MST Relative to P&P.....	90
4.3 Percentage of Items Exposed to Different Numbers of Examinees in Each Test Design.....	93
4.4 Number of Items Available and Number of Items Exposed	95
4.5 Conditional Exposure Rates for Each Test Design	97

LIST OF FIGURES

Figure		
2.1	Example of a 36-Item Multi-Stage Test with 18 Items at Each Stage	27
3.1	Subpool Information.....	59
3.2	36-Item Two-Stage Test with Three Subtests at the Second Stage.....	60
3.3	Example of a 36-Item Multi-Stage Test with 18 Items at Each Stage	61
3.4	Frequency Distribution of b Parameters in the Pool	63
4.1	RMSEs of All 14 Test Designs	76
4.2	Frequency Distribution of b Parameters in the Pool	77
4.3	RMSEs of <u>Two</u> -Stage Tests with <u>Three</u> Subtests in the Second Stage MST (I-III).....	78
4.4	RMSEs of <u>Three</u> -Stage Tests with <u>Three</u> Subtests in the Second and Third Stages (MST IV-VI)	79
4.5	Comparison of Two- and Three-Stage Tests with <u>Three Subtests</u> in Second and Third Stages	79
4.6	RMSEs of <u>Two</u> -Stage Tests with <u>Five</u> Subtests in the Second Stage (MST VII-IX)	80
4.7	Comparison of <u>Two-Stage Tests</u> with Three or Five Subtests in the Second Stage	80
4.8	RMSEs of <u>Three</u> -Stage Tests with <u>Five</u> Subtests in the Second Stage (MST IX-XII)	81
4.9	Comparison of Two- and Three-Stage Tests with <u>Five Subtests</u> in the Second and Third Stages	82
4.10	Comparison of <u>Three-Stage Tests</u> with Three or Five Subtests in the Second and Third Stages	82
4.11	Bias of All 14 Test Designs.....	84

4.12	Efficiency of MST I-III Relative to CAT and P&P	86
4.13	Efficiency of MST IV-VI Relative to CAT and P&P.....	87
4.14	Efficiency of MST VII-IX Relative to CAT and P&P	87
4.15	Efficiency of MST X-XII Relative to CAT and P&P.....	88

CHAPTER 1

INTRODUCTION

1.1 Background

Traditionally, testing organizations and credentialing/licensing agencies have used paper-and-pencil (P&P) tests to measure an examinee's ability, knowledge, skill, or competency in a particular domain. Today, however, with the efficiency and affordability of computers and the potential advantages that may be derived from testing using computers, many testing organizations and credentialing/licensing agencies are considering the use of computer-based testing (CBT). Advantages of CBT include increased validity through the computer's capacity to support new and innovative item formats, the potential to improve test security, and its ability to obtain additional information for assessing proficiency from the speed of response. Other advantages include economy of paper, improved data collection and pretesting of items, year-round testing, the convenience and flexibility of individual scheduling, immediate feedback to examinees which can be beneficial for diagnostic purposes, and faster score reporting services (Wainer, Dorans, Flaugher, Green, Mislevy, Steinberg, & Thissen, 1990).

Of these advantages, the greatest advantage of CBT appears to be the potential to increase test score validity through the computer's capacity to support new and innovative item formats. However, along with many new and innovative item formats comes the need for polytomous scoring models that allow one to assess partial knowledge on more realistic tasks than are measured by typical P&P multiple-choice tests (e.g., computer-based case simulations in medical licensing; Clyman, Melnick, & Clauser, 1995).

Fortunately, as with P&P testing, CBT supports the use of polytomous item response models to score tests.

In moving to CBT, some testing programs have simply sought to transfer their existing P&P test onto a computer and deliver the test via the computer. This type of CBT is referred to as linear CBT. As with conventional P&P tests, all examinees who take a linear computer-based test are administered the same fixed set of items. Usually, after a prescribed amount of time, a new linear computer-based test is introduced. While all of the advantages of CBT are present in linear CBT, if one is to offer year-round testing with the convenience and flexibility of individual scheduling, there arises a security problem in that examinees could see the same items within a prescribed interval of time. Examinees could then share items with future examinees taking the test within the same interval of time, possibly causing future examinees' test scores to be invalid (Patsula & Steffen, 1997).

Although simply administering tests via the computer improves test security by minimizing the chances of a test booklet "leaking" during shipping or the chances of a proctor adjusting score sheets, the computer can enhance test security further by minimizing the benefit of foreknowledge of items. Given a pool of precalibrated items, the computer can design strictly parallel forms and administer the different forms to examinees at random. This type of CBT is known as randomly equivalent forms CBT or linear-on-the-fly testing (LOFT).

While randomly equivalent forms CBT improves upon linear CBT, one criticism of both approaches is that one is not using the computer to its full capacity. "The computer can do more than simply administer a predetermined set of test items. Given a pool of

precalibrated items to choose from, the computer can design a different test for each examinee.” (Lord, 1980, p. 150). The advantage of designing a test for each examinee is that the test can be geared to the examinee’s ability so items that are either too easy or too difficult for the examinee are not included in the test. An examinee’s ability is measured most effectively this way. This type of testing is known as adaptive testing.

In adaptive testing, items are tailored to an individual examinee’s ability level through branching to more difficult items following correct responses and branching to easier items following incorrect responses (Lord, 1970). Consequently, each examinee potentially sees a different form of the test. Although adaptive tests can be administered both by P&P (Lord, 1971a, 1971b) and computer (Lord, 1970), today, they are predominantly computerized. Thus, the term computerized adaptive testing (CAT) is used.

A distinct advantage of CAT is that it offers the potential of a shorter test since items that are too easy or too difficult for the examinee are not administered, unless of course an item is needed to satisfy some content specification or to avoid overexposure of another item. This “tailoring” of items to an examinee’s ability level leads to adaptive tests that are often more efficient than conventional P&P tests (Lord, 1980; Weiss, 1982), typically requiring examinees to answer about half as many items to attain an equivalent level of precision (Green, 1983; Olsen, Maynes, Slawson, & Ho, 1986; Schnipke & Reese, 1997).

Because of the many advantages associated with CAT, a number of organizations have moved from P&P testing to CAT. The first known organization to commit to large-scale CAT was the U.S. Armed Services, who currently use a CAT version the Armed

Services Vocational Aptitude Battery (ASVAB) for the psychological examination of recruits to assign them to training school and job specialties. Examples of other programs that use CAT are the National Council of State Boards of Nursing (National Council Licensure Examination – NCLEX), the Graduate Record Examination Board (Graduate Record Examination – GRE), and the Graduate Management Admissions Council (Graduate Management Admissions Test – GMAT).

Although there are many advantages associated with CAT, there are four main criticisms. First, examinees taking a computerized adaptive test are typically not permitted to review their answers to previous questions. Not surprisingly, examinees report this as the greatest criticism of CAT (Lunz, Bergstrom, & Wright 1992). Although research has shown that there are ways to allow item review in CAT and still obtain accurate estimates of ability (Lunz et al., 1992; Stocking, 1996; Stone & Lunz, 1994; Wise, 1996), item review has yet to be implemented in practice.

A second criticism of CAT is that the number of items exposed in a computerized adaptive test is quite high (Luecht, Nungester, & Hadadi, 1996). While there are exposure controls built into CAT algorithms, the purpose of the controls tends to be to reduce item exposure rates (i.e., the number of people seeing an item) rather than to reduce the number of items exposed (Stocking, 1993; Stocking & Lewis, 1998). Exposing many items, regardless of how many examinees see the items, can affect the accuracy and validity of test scores if future examinees gain access to exposed items prior to testing.

Third, the fact that it is sometimes difficult to content balance a computerized adaptive test while still maximizing reliability and satisfying exposure controls is another

criticism of CAT (Luecht et al., 1996). Content balancing a test poses a considerable challenge in CAT, when there are many content constraints to satisfy. In building a P&P test form, test specialists focus primarily on maximizing reliability while content balancing the test and item exposure is not an issue since items are disclosed after each test is administered. In CAT, item pools are not disclosed after each administration and so test specialists now need to take into account item exposure. This places considerably more demands on the item pool. In fact, though computerized adaptive tests tend to be shorter than P&P tests, overall, more items are required to maintain an operational CAT program than a P&P testing program.

Finally, in CAT, millions of different test forms are possible from a single item bank and it is, therefore, not feasible for test specialists or committees to review every test form for quality assurance purposes (Luecht & Nungester, 1998). Although a CAT item selection algorithm can provide some quality assurance by the inclusion of specified content and other categorical constraints in the item selection algorithm (Stocking & Swanson, 1993), the algorithm is limited to what can be coded numerically about each item (item format, content specifications, etc).

These criticisms of CAT are serious and have hampered the implementation of CAT by many testing agencies. An alternative to CAT that eliminates some of the criticisms of CAT is multi-stage testing (MST) or what some know as testlet-based testing. While linear CBT and CAT represent the two extremes of CBT, one variation of CBT that lies in the middle of the continuum is computerized MST. Computerized MST is a compromise between CBT and CAT and is, in fact, a special case of CAT that allows for item review, reduces the number of items exposed, balances content regardless of the

number of constraints, makes the implementation of quality assurance more feasible, and still maintains all of the advantages of CBT. Furthermore, a multi-stage test can be administered either by P&P or by computer.

In MST, there is partial adaptation of the test to individual examinees. However, rather than adapting the test to individuals item by item as in CAT, the test adapts to the examinee in stages or sets of items. In MST, all examinees are administered a common set of items known as a routing or stage-one test. Depending on examinee performance, the examinee is routed to one of several alternative second-stage tests, each of which consists of a fixed set of items and differs on average difficulty. Depending upon examinee performance on the second-stage test, he or she is routed to one of several alternative third-stage tests. This process continues depending on the number of stages in the MST procedure. The number of stages and the number of subtests per stage, among other factors, vary between different testing programs that utilize MST. However, what is most commonly found among organizations that employ MST is a two-stage testing procedure with three subtests contained in the second stage (Luecht & Nungester, 1998; Rock, 1996; Rock et al., 1995).

While MST appears to eliminate some of the common criticisms of CAT, inherent in MST procedures are two drawbacks: the potential decrease in accuracy of ability estimation and a likely loss of efficiency relative to CAT (Loyd, 1984; Kim & Plake, 1993; Luecht et al., 1996; Schnipke & Reese, 1997). Nonetheless, in weighing the advantages and disadvantages of MST, many testing programs have either implemented MST (e.g., National Board of Medical Examiners) or are considering the implementation of MST (e.g., Law School Admission Council).

With increased interest in MST, more research is needed in MST. To date, there appears to be less research in the area of MST than CAT, though the history of testing might suggest otherwise, as MST was introduced in the mid 1960s prior to CAT. MST is inundated with design issues. Examples of such issues are the number of stages to have or the number of subtests to have for any particular stage, and the number of items to include in each subtest to achieve accurate measurement (Luecht et al., 1996; Luecht & Nungester, 1998). While some researchers have studied MST design issues (e.g., Lord, 1980; Loyd, 1984; Kim & Plake, 1993; Luecht et al., 1996), several questions remain to be addressed. Since numerous testing programs are currently deciding whether to implement MST or CAT, it seems important to compare the two forms of testing.

1.2 Multi-Stage Testing

In its earliest form, multi-stage tests were known as sequential or flexilevel tests (Cronbach & Gleser, 1965; Lord, 1971a, 1971b) and were primarily two-stage tests that were administered by P&P. Cronbach and Gleser (1965) introduced two-stage P&P testing using a decision theory approach. Their purpose for testing was to pass or fail an examinee. Since their primary interest was to classify an examinee, they used a sequential approach of administering a second-stage test to only borderline examinees; those near, but not quite reaching, the passing score. The advantage of this was that one could save time by not testing people further if the first stage was definitive as to whether they were going to pass or fail.

In contrast to Cronbach and Gleser's purpose for two-stage testing, Lord's purpose for two-stage testing was to improve measurement of an examinee's ability,

rather than simply to classify examinees (Lord, 1971b). Since conventional P&P tests were judged already to measure the typical examinee well (an examinee found where the majority of people are found; an examinee in the middle of the ability range), Lord's intent with two-stage testing was to improve measurement at the extreme ability levels by matching the second-stage test to the ability level of the examinee. As in CAT, such tailoring of a test to an examinee's ability level would produce the advantage of a potentially shorter test. Furthermore, it would avoid the undesirable or demoralizing effect of a test being excessively difficult or easy for an examinee (Lord, 1971b).

The potential of shorter tests and avoiding demoralization of examinees, coupled with the advantages of item review, exposing fewer items, content balancing, and a fixed number of test forms that test specialists can review for quality assurance purposes are the primary reasons why testing organizations are turning to MST. Furthermore, with the advantages associated with CBT, organizations are considering computerized MST.

More research is needed in the area of MST. While several researchers (Lord, 1980; Loyd, 1984; Kim & Plake, 1993; Luecht et al, 1996; Luecht & Nungester, 1998) have contributed to our knowledge in terms of what factors to consider in designing a multi-stage test, several important issues surrounding MST remain to be resolved. For example, given limitations on the number of items in a test, how might a multi-stage test be designed to maximize measurement accuracy or minimize classification errors? Another question of interest may be principles for utilization of the item bank. CAT, in principle, exposes more items than does MST (Luecht et al., 1996). Does MST offer a better way to optimize the use of your item bank?

1.3 Statement of the Problem

In recent years, interest has grown in the assessing of proficiency via MST rather than P&P testing or other CBT. Medical licensing exams are important examples of such assessments. Several issues in MST need to be addressed. In general, the following factors need to be addressed in designing a multi-stage test (Lord, 1980):

1. Total number of items. In general, as with any test, increasing the number of items in a multi-stage test, by adding items of comparable quality to the existing items, will increase the reliability of the test and, hence, improve measurement (Loyd, 1984; Kim & Plake, 1993). With regard to test length, the primary issue in MST is: Given a test of fixed length, how many items should reside in each stage?
2. Number of items in the routing test. The number of items in a routing test can have a profound effect on the routing of an examinee to the second-stage test, and possibly to other stages as well, and can ultimately affect the accuracy of ability estimation. With the total number of items fixed, if the routing test is too long, there will be few items in the second-stage test and, in essence, the test will no longer be adaptive. Conversely, if the routing test is too short there will be poor allocation to the second-stage test and possibly poor measurement of ability (Lord, 1980). Both Loyd (1984) and Kim and Plake (1993) varied the number of items in the routing test and found that the longer routing test was superior. However, neither kept total test length constant. The question now becomes, with the total number of items in a test fixed, what portion of items should be placed in the routing test?

3. Difficulty level of the routing test. There is a consensus among practitioners that the routing test should be of moderate difficulty. However, a question of concern is: What type of distribution should the difficulty of the items in the routing test have? Alternatively, what type of distribution should the information of the items in the routing test have?

A variable of considerable importance relative to the difficulty level or information generated by the routing test is the expected distribution of examinee ability. Kim and Plake (1993) found that given a rectangular distribution of examinee ability, a rectangular distribution of item difficulty produced more accurate ability estimates than a peaked distribution of difficulties. Kim and Plake (1993) confined their research to the case of a rectangular distribution of examinee ability. Other distributions have yet to be investigated. One might hypothesize that given a peaked distribution of examinee ability, a routing test with a peaked distribution of item difficulties or information would be highly desirable. It is common for a testing organization to know the expected ability distribution of the population with which they are dealing. This information can be important in designing a multi-stage test, but to date has not been investigated.

4. Routing of the examinees. It is important that the difficulty levels of the second-stage test match the ability levels of the examinees allocated to them, as determined by the routing of the examinees. Likewise, the difficulty levels of the third-stage tests should match the ability levels of the examinees allocated to them. Determining how to route examinees is an important issue, but it has received little attention in the literature.

5. Number of stages. While some researchers have investigated two- and four-stage tests (Luecht et al., 1996; Schnipke & Reese, 1997), it appears that no studies have systematically compared the accuracy of ability estimates obtained from varying the number of stages.
6. Number of subtests per stage. Several studies have shown that the number of second-stage tests influences the accuracy of measurement (e.g., Lord, 1971b; Kim & Plake, 1993). However, there have been no studies that have investigated the effect of the number of subtests at higher stages on ability estimation. Luecht, Nungester, and Hadadi (1996) and Luecht and Nungester (1998) suggest that more research is needed on the effects of the number of subtests per stage on the accuracy of ability estimation.
7. Difficulty levels of the subtests. Lord (1971b) found that if the difficulty levels of the second-stage tests are too close to the level of the difficulty level of the routing test, poor measurement is obtained at the extreme ability levels; if the difficulty levels are too extreme, there is poor measurement where the ability level of the examinee was too near the difficulty level of the routing test. In addition to the difficulty levels of the routing test, a concern is the distribution of difficulty levels of the subtests. One hypothesizes that the distribution of difficulty levels of the second-stage tests should probably be peaked according to the cut-points on the subtests for routing examinees to a third-stage test. Similarly, the distribution of difficulty levels of the third-stage tests should be peaked according to cut-points on the second-stage test.

At this point, one may foresee the problem that can arise from peaking item difficulty of a subtest according to cut-points on the previous stage test and peaking it according to cut-points on subsequent stage tests. It would lead to a multi-modal distribution of item difficulties, which would be the same as having a uniform distribution of item difficulties. Hence, a uniform distribution of item difficulties is desirable.

Alternatively, one may want to consider the distribution of information of the subtests rather than distribution of item difficulties. In this way, information obtained from item discrimination and guessing are taken into account along with information from the item difficulty.

8. Overlap of difficulties of subtests at the same stage. If, for example, the routing test is fairly long, then overlap of difficulties in subtests at the next stage may not be very important. However, the overlap may be integral if the routing test is short since there will be less confidence in the ability estimates obtained from the first stage (Lord, 1980). There appear to be no studies in the literature that have investigated the effect of the amount of overlap on the accuracy of ability estimation.
9. Method of scoring the total test. Today, it is most common to use maximum likelihood scoring procedures for estimating ability (Loyd, 1984; Kim & Plake, 1993; Schnipke & Reese, 1997). The reason for such usage is twofold: one, it is easier to implement maximum likelihood scoring procedures than expected a posteriori procedures, and two, there is little difference between the two in the estimation of ability (Luecht et al., 1996).

1.4 Purpose of the Study

The topic of this study was the design and evaluation of multi-stage tests. Section 1.3 listed many variables that enter into the design of a multi-stage test. For any given multi-stage test, some variables may be fixed, while others may vary. The purpose of this study was to compare the accuracy of ability estimates produced by MST and CAT while keeping some variables fixed and varying others. In this study, item information rather than item difficulty was considered since item information encompasses more information about an item than does item difficulty alone.

The variables that were held constant were the total number of items in the test, the methods for determining the distribution of information of the subtests, the amount of overlap of information between subtests at the same stage, the method for routing examinees, and the method utilized in scoring the total test. In MST, total test length is usually fixed and, thus, was fixed in this study. Secondly, the same empirical method for determining the distribution and amount of overlap of information for each subtest was used. In addition, the method for routing examinees was not varied. A method that routes examinees to the stage test that provides maximum information given the examinee's ability estimate was used (R. Luecht, personal communication, November 16, 1998). Finally, since little difference was found between ability estimates obtained from maximum likelihood and expected a posterior estimation (Luecht et al., 1996), method of scoring total test was not manipulated.

In summary, only the following factors were systematically varied:

- a. number of stages,

- b. number of subtests per stage, and
- c. number of items per subtest.

The above noted factors appear to be the most salient factors in MST design and appear to be factors that researchers believe deserve more attention in designing multi-stage tests (Luecht et al., 1996; Luecht & Nungester, 1998).

The primary question of interest was: Given a fixed total test length and controlling for item exposure, how many stages and how many subtests per stage should there be to maximize measurement precision? Furthermore, how many items should be assigned to each subtest? Should there be more in the routing test or should more items be allocated to the higher stage tests? A secondary question of interest concerned conditional item exposure rates and the number of items exposed by CAT and the different MST designs.

1.5 Significance of Problem

There is considerable evidence that demonstrates that CAT and MST are viable frameworks for testing. With many testing organizations looking to move toward CAT or MST, it was important to ascertain which framework functions superiorly in different situations in terms of measurement accuracy and item exposure. What was needed was a systematic comparison of the different testing procedures under various realistic testing conditions. This dissertation addressed the paramount problems of the increase or decrease in accuracy of ability estimation and item exposure rates in using MST rather than CAT. The expectation, too, was that some guidelines would result that would influence multi-stage test design.

CHAPTER 2

REVIEW OF LITERATURE

In this chapter, the importance of item response theory in computerized adaptive testing is reviewed and computerized adaptive testing designs and multi-stage testing designs and issues are discussed.

2.1 The Importance of Item Response Theory in Computerized Adaptive Testing

In 1970, Lord introduced the notion of computerized adaptive testing (CAT) with the use of item response theory (IRT). With computerized multi-stage testing (MST) being a special case of CAT, IRT also plays a central role in MST. When the fit between an IRT model and test data of interest is satisfactory, IRT models are said to provide invariant item and ability parameters (Lord, 1952). “This [invariance] property implies that the parameters that characterize an item do not depend on the ability distribution of examinees [sample-free item parameters] and the parameter that characterizes an examinee does not depend on the set of test items [test-free ability parameters]” (Hambleton, Swaminathan & Rogers, 1991, p. 18). Sample-free item parameters and test-free ability parameters are further explicated, respectively.

IRT’s property of sample-free item parameters allows one to use an IRT model to calibrate items in a bank to a common scale, even when different groups of examinees responded to the items. Equivalence of examinee samples is achieved via statistical adjustments obtained from a linking design. In turn, a CAT item selection algorithm can select items from the item bank that provide the most information (based on the item

parameters that are on the same scale) about an examinee's ability given the examinee's current ability estimate from his or her responses to previous items. Information for item i given examinee j with ability θ is expressed as:

$$I_i(\theta_j) = \frac{2.89\alpha_i^2(1-c_i)}{(c_i + e^{1.7\alpha_i(\theta_j-b_i)})(1 + e^{-1.7\alpha_i(\theta_j-b_i)})^2},$$

where α_i , b_i , and c_i are the IRT discrimination, difficulty, and pseudo-guessing parameters, respectively, for item i .

Likewise, in MST, an examinee is routed to a stage test that provides the most information about the examinee's ability given the examinee's ability estimate obtained from the previous stage test. The information obtained from a multi-stage test for examinee j with ability θ is simply the sum of information of each item in the stage test:

$$T(\theta_j) = \sum_{i=1}^n I_i(\theta_j),$$

where n is the number of items in the stage test. It is well known that item or test information is maximized when item or test difficulty is close to the examinee's ability (Hambleton et al., 1991). For this reason, item and test selection procedures in CAT or MST are most commonly based on maximum item or test information.

The test-free ability parameter property allows one to compare examinees' ability estimates even when they are based on different tests of varying difficulty. Implicit in the ability estimate is the difficulty of the items. This is crucial to the success of both CAT and MST, since in both types of tests different examinees are administered different tests.

Today, CAT and MST rest on IRT's invariance property and IRT's information function. In the following sections, CAT and MST are described in more detail, respectively.

2.2 Computerized Adaptive Testing

As is evident in the measurement literature of the past ten years, the use of computerized adaptive testing (CAT) by testing organizations and credentialing/licensing agencies to measure an examinee's ability, knowledge, skill, or competency in a particular domain has become increasingly prominent. This prominence can be attributed to the many advantages that CAT has to offer to both test takers and test developers. Such advantages are derived from both the computer delivery of the test, as well as the adaptive nature of the test.

In adaptive testing, examinees are administered items based on their responses to previous items. Items in an adaptive test are tailored to an individual examinee's ability level through branching to more difficult items following correct responses and branching to easier items following incorrect responses. Although adaptive tests can be administered by paper-and-pencil (P&P) or computer, today, they are predominantly computerized. Thus, the term computerized adaptive testing or CAT is used.

Lord introduced the notion of CAT in 1970. In 1970, however, the idea of CAT was purely theoretical, as it was not "convenient to use computers to administer achievement tests" (Lord, 1980, p.150) due to the cost and scarcity of computers. Today, however, with the widespread availability and use of computers, CAT is feasible. In fact, there are many testing organizations that utilize CAT.

2.2.1 Examples of CAT

The first well-known organization to commit to large-scale CAT was the military services of the United States. Since 1984, the U.S. military has had a computerized adaptive screening test, which is an initial screening test used by army recruiters to determine if a prospect is suitable to send to a Military Entrance Processing Station (MEPS) for full-scale testing with the Armed Services Vocational Aptitude Battery (ASVAB). The ASVAB is used for the psychological examination of recruits to assign them to training school and job specialties (Sands, Waters, & McBride, 1997). The CAT version of the ASVAB has existed since 1992, however, operational use of the CAT-ASVAB only began in 1997. Today, the CAT-ASVAB is administered at 60 MEPS nationwide.

Examples of other organizations that use CAT are the National Council of State Boards of Nursing which develops the National Council Licensure Examination (NCLEX), the Graduate Record Examination Board that develop the Graduate Record Examination (GRE) General Test, and the Graduate Management Admission Council which sponsors the Graduate Management Admission Test (GMAT). As of 1992, anyone aspiring to be licensed as a registered or public nurse is required to take the CAT version of the NCLEX (M. Potenza, personal communication, March 8, 1998). Effective in Fall of 1993, anyone taking the GRE General Test had the option to take a CAT or P&P version of the test (Schaeffer, Steffen, Golub-Smith, Mills & Durso, 1995). The GRE General Test measures verbal, quantitative, and analytical skills and is designed to help predict a student's performance in the first year of graduate school. Finally, unlike the

GRE, which offers test takers the option to take the test by P&P or by CAT, as of October 1997, people who wish to take the GMAT must take the GMAT-CAT, with the exception of people in countries where there are no testing centers (Graduate Management Admission Council, 1997). Similar to the GRE, the GMAT measures language, quantitative, and analytical writing skills and is designed to help predict a student's potential academic performance in the first year of graduate management school.

There are many advantages of CAT to both test takers and test developers. However, inherent in CAT are some disadvantages and challenges that need to be addressed. In the sections to follow, these issues are explained and explored.

2.2.2 Advantages of CAT

From an examinee's viewpoint, adaptive testing offers the advantage of a potentially shorter test since items that are too easy or too difficult for the examinee are not administered, unless a particular item is needed to satisfy some content specification or to avoid overexposure of another item. This tailoring of items to an examinee's ability level can lead to an adaptive test that is often more efficient than conventional P&P tests, typically requiring examinees to answer about half as many items to attain an equivalent level of precision (Green, 1983; Olsen, Maynes, Slawson, & Ho, 1986; Schnipke & Reese, 1997; Weiss, 1982).

Other advantages of CAT to test takers are derived from the computerized delivery of the test. In most cases, computerized adaptive tests are administered in testing centers that are located in major cities and towns. Each testing center is usually equipped with a minimum of five to six computers and is open for testing Monday through Friday 9

a.m. to 5 p.m., except for holidays; some testing centers even offer appointments on Saturdays. This offers to test takers the advantages of year-round (daily) testing and the convenience and flexibility of individual scheduling (Green, 1983; Wainer, Dorans, Flaughner, Green, Mislevy, Steinberg, & Thissen, 1990). Each test taker is free to choose what day and time (morning or afternoon) he or she would like to take the test. Some argue that this really is not an advantage because if test takers were given the choice of when to test, they would all choose to test Saturday morning and testing centers would not be able to accommodate them all. The result would be that the test taker really does not have the convenience or flexibility of scheduling. However, a counter-argument could be made that a large percentage of the test taker population consists of students who would most likely NOT want to schedule a test for Saturday morning. Thus, the convenience and flexibility of scheduling remains a potential advantage to test takers.

Finally, other advantages to test takers derived from the computerized nature of CAT are the immediate score feedback and faster score reporting to institutions (Green, 1983; Wainer et al., 1990). Since all items in a CAT item pool are precalibrated, the computer can produce a score immediately after testing and hence there is the potential for faster score reporting to schools and credentialing agencies.

In summary, advantages of CAT to test takers are potentially shorter tests, year-round testing, the convenience and flexibility of individual scheduling, immediate knowledge of scores, and expedited score reporting services.

Advantages of CAT to test developers include improved test reliability and validity through improved data collection, the potential of improved test security, tightened

controls on cheating, cost savings with regard to printing and shipping, and the opportunity to support new measurement (Green, 1983; Lord, 1980; Wainer et al., 1990).

The opportunity for improved data collection and greater control of cheating are derived from the computerized nature of CAT. First, data no longer need to be collected and then shipped and scanned, but instead are simply transmitted electronically. Electronic transmission of data eliminates errors in scanning and decreases the cost and need for printing and shipping test booklets and answer sheets. This further makes CAT an environmentally friendly venture, which in today's society is highly valued. Electronic transmission of data also improves test security by minimizing the chances of a test booklet "leaking" during shipping or the chances of a proctor adjusting score sheets. As a result of improved data collection, and, thus, bolstered security, there is further assurance that examinees' scores are valid (Federation of State Medical Boards of the United States, Inc. & National Board of Medical Examiners, 1998). Secondly, since the proctor to test taker ratio is very low in CAT, there is a better opportunity to control cheating from a neighboring examinee, from notes, or from examinees misrepresenting themselves. Once again, this increases confidence in the validity of test scores.

Although the aforementioned advantages of CAT to test developers are important because they can lead to increased validity in test scores, the biggest advantage of CAT to test developers appears to be the opportunity of the computer to support new media for measurement purposes, such as video clips and simulations. Currently, the National Board of Medical Examiners is developing an innovative test in which a patient-care environment is simulated. The examinee is presented a scenario on the computer and responds by ordering certain actions, such as taking a medical history, ordering certain tests, or

prescribing certain medications. The patient's condition adapts/changes (i.e., gets better or worse) depending on the examinee's actions. It is argued that tests such as these more accurately reflect what the examinee will need to do in practice and, thus, increase the validity of test scores (Clyman, Melnick, & Clauser, 1995).

In summary, for test developers, CAT offers the advantages of improved test reliability and validity through improved data collection and test security, a better opportunity to control cheating, and the opportunity to support new measurement. It also offers a mean to curb costs associated with printing and shipping test booklets and answer sheets.

2.2.3 Disadvantages to CAT

Although the abovementioned advantages to test takers and developers remain as potential benefits to CAT, they are often overshadowed by many practical concerns and challenges that researchers have come to learn through operational CAT programs. Challenges to be confronted by test developers in implementing CAT include practical issues such as selecting the first item, administering items belonging to sets, controlling item exposure and overlap, providing item review to test takers, selecting a stopping rule for variable length tests, scoring adaptive tests, allowing for incomplete tests, writing enough items to provide appropriate tailoring of the test and to maintain the security of items, developing and maintaining CAT pools, developing new item types and its associated cost, setting up testing centers, and complying with disclosure requirements (Mills & Stocking, 1995).

To test takers, the single greatest criticism of CAT is not providing item review to examinees. In general, examinees are not permitted to review or skip items in a computerized adaptive test. That is to say, examinees can not go back in the test and change their answers or “check” their answers; they usually can not skip questions either. One concern has been that reviewing and altering item responses or skipping items “may change the estimate of examinee ability such that the sequence of items will become poorly targeted and precision will be lost” (p. 34, Lunz, Bergstrom, & Wright, 1992). Test developers believe that not allowing examinees to review items provides test developers optimal psychometric control over the test (Lunz & Bergstrom, 1994). Not surprisingly, examinees feel at a disadvantage when they cannot review and alter their responses.

While there have been many researchers who have studied the effects of reviewing and altering responses on examinees’ ability estimates in a computerized adaptive test, the results are inconclusive (Wise, Barns, Harvey, & Plake, 1989; Lunz et al., 1992; Lunz & Bergstrom, 1994). Wise, Barns, Harvey, and Plake (1989) found no significant differences in the mean scores of examinees in review and non-review conditions, while Lunz, Bergstrom, and Wright (1992) and Lunz and Bergstrom (1994) found that examinees who were allowed to review performed significantly better on average than the examinees who were not allowed to review. However, within the review condition of both studies (Lunz et al., 1992; Lunz & Bergstrom, 1994), there were no significant differences in mean ability estimates between students who used the review option and those who did not use the review option. This suggests that examinees simply knowing that they have the option to review items is helpful; their actually reviewing and altering responses has no significant impact on examinee performance.

In addition to examining the effects of review on ability estimates, Lunz, Bergstrom, and Wright (1992) also studied its effect on the efficiency of CAT. They found that the average efficiency of the test decreased by only 1% after review and that on average, the information loss could be recovered by the addition of less than one item.

Based on the results of these studies (Wise et al., 1989; Lunz et al., 1992; Lunz & Bergstrom, 1994), a solution to allowing item review may include a variable length computerized adaptive test that continues administering items after review until ability is estimated with a certain amount of predetermined precision (e.g., $\pm 1.96\text{SEM}$). However, this approach has item exposure implications – more items will be exposed. Another concern that these researchers (Wise et al., 1989; Lunz et al., 1992; Lunz & Bergstrom, 1994) do not take into account in supporting item review in CAT is the scenario where examinees purposely answer all items wrong to obtain easier items and then correct previous items to receive a new perfect raw score that can lead to a very high ability estimate.

For test developers, two advantages of CAT to test takers are actually disadvantages to test developers. Initially, one might think that the reduction in items on any given test would also be an advantage to test developers. One would think that if examinees require shorter tests, then test developers would not need to develop as many items. However, this is not true. Examinees span the full ability spectrum and so medium and difficult tests are needed along with hard and easy tests to match all examinee ability levels. In addition, the administration of “on-demand” tests discloses items on a daily basis and large item banks therefore are needed for this reason as well. Despite the fact that the adaptive nature of CAT allows for the potential of shorter tests for examinees, this

does not carry over to test developers in allowing them to develop fewer items.

Furthermore, the advantage of daily testing for test takers does not provide an advantage in economy to test developers. To minimize item exposure rates in daily testing and, hence, to increase test security, test developers need to develop larger items banks than are typically needed for P&P testing.

Finally, researchers investigating multi-stage testing (MST) present three criticisms of CAT as it relates to test developers. First, since each examinee may be theoretically administered a different test form, millions of different test forms are possible from a single item bank, making it unfeasible for test specialists or committees to review every test form for quality assurance purposes (Luecht & Nungester, 1998). Although a CAT item selection algorithm can provide some quality assurance by the inclusion of specified content and other categorical constraints in the item selection algorithm (Stocking & Swanson, 1993), the algorithm is limited to what can be coded numerically about each item (item format, content specifications, etc).

In defense of CAT, one must keep in mind that human review of final test forms may not be equally valued in all testing programs. Human review can be very costly in terms of time and efficiency. On the other hand, full computer automation is fast and efficient (van der Linden & Boekkooi-Timminga, 1989). The amount of quality assurance needed by humans may vary depending on the purpose of the test, on test development policy, and financial implications. High-stakes licensing programs (e.g., medical licensure programs) would most likely hesitate to rely solely on a computer algorithm to assure quality of a test form for fear of a computer algorithm missing “something” in such a high-stakes licensure test.

A second criticism of CAT is that more items are exposed in CAT than MST (Luecht et al., 1996). While there are exposure controls built into CAT algorithms, the purpose of the controls tends to be to reduce item exposure rates (i.e., the number of people seeing an item) rather than to reduce the number of items exposed (Stocking, 1993; Stocking & Lewis, 1998). Exposing many items, regardless of how many examinees see the items, can affect the accuracy and validity of test scores if future examinees gain access to exposed items prior to testing.

Finally, according to Luecht, Nungester, and Hadadi (1996), it is sometimes difficult to content balance a computerized adaptive test while still maximizing reliability or minimizing decision errors and satisfying exposure controls. This is a considerable challenge in CAT, particularly if there are a multitude of content constraints to satisfy. Nonetheless, CAT programs such as the GRE and GMAT are managing to overcome this obstacle (personal communication, M. Steffen, March 11, 1998).

In summary, the criticism of CAT by examinees is that typically they are not permitted to review and alter their responses. Criticisms of CAT by MST researchers are the lack of opportunity for test specialist or committee review of final test forms, the fact that item exposure can be large, and that it is often difficult to balance content in CAT. In addition, the nature of administering adaptive tests on a daily basis requires larger items banks than are typically needed for P&P testing. All of these criticisms have some validity and may be more problematic in some testing programs than others. An alternative to CAT, which addresses some of these issues, is MST.

2.3 Multi-Stage Testing

In the literature, multi-stage testing (MST) chronologically followed CAT (Lord, 1970, 1971b). Without the widespread use and availability of computers, CAT was not feasible in practice. MST was introduced as the “poor man’s version” of CAT – a P&P version of adaptive testing. Today, however, with the power and extensive availability of computers, testing organizations that are considering MST are considering producing, administering, and scoring multi-stage tests via computer.

In MST, there is some adaptation of the test to individual examinee ability level. However, rather than adapting the test to individuals item by item as in CAT, the test adapts to individuals in stages represented by groups of items (see Figure 2.1).

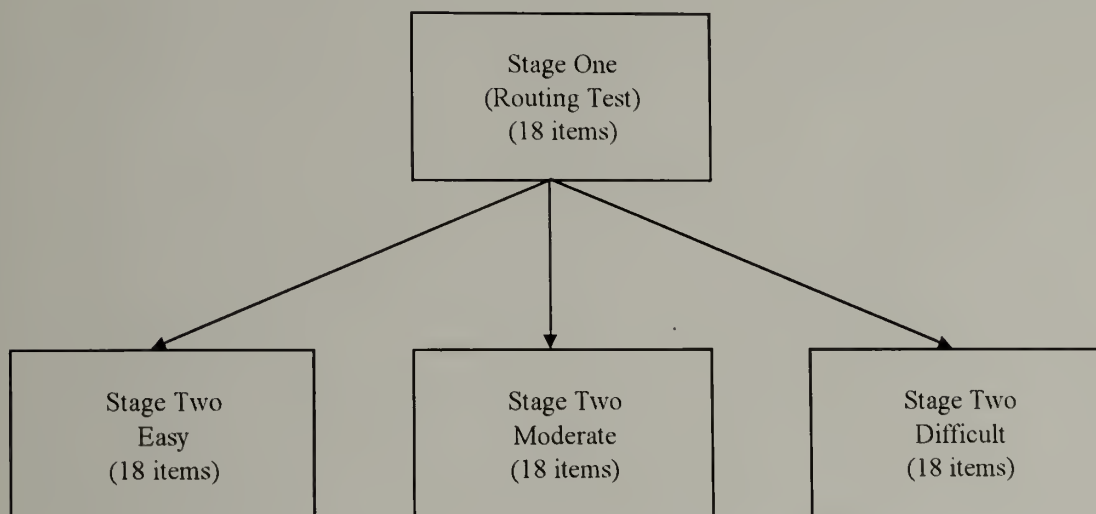


Figure 2.1. Example of a 36-Item Multi-Stage Test with 18 Items at Each Stage

In MST, all examinees are first administered a common set of fixed items known as a routing test. Depending on how an examinee performs on the routing test, he or she is routed to one of several alternative second-stage tests, each of which consists of a fixed

set of items and differs in average difficulty level. Depending on how the examinee scores on the second-stage test, he or she is routed to one of several alternative third-stage tests. This process continues depending on the number of stages in the MST procedure.

The number of stages and the number of subtests per stage, among other factors, vary between different testing programs that utilize MST. However, what is most commonly found among organizations that employ MST today is two-stage testing with three second-stage tests (Rock, Pollack, & Quinn, 1995; Luecht et al., 1996; Rock, 1996).

2.3.1 Examples of Multi-Stage Testing

Described below are two studies and one testing program that utilize MST. The two studies used P&P multi-stage tests, whereas the testing program used a computerized multi-stage test. The first study described is the National Education Longitudinal Study of 1988 (NELS:88; Rock et al., 1995). The NELS:88 was “designed to monitor the transition of a national sample of young adults as they progress[ed] from eighth grade to high school and then on to postsecondary education and/or work” (Rock, et al., 1995, p. 1).

To minimize floor effects in eighth grade and ceiling effects in twelfth grade, to make the assessment of gain more accurate, and to keep testing time to a minimum, Rock and his colleagues (1995) employed MST as opposed to a very long test with many easy items, as well as many difficult items. In particular, they used two-stage testing with three stage-two tests in mathematics and two stage-two tests in reading at both grades 10 and 12. Students were routed to a “stage-two” test based on their ability estimate from the previous test. For example, depending on how a student performed in grade 10, he or she was routed to an appropriate grade 12 stage-two test. It was found that the two-stage

procedure increased the accuracy of the measurement, and, when used in combination with Bayesian item parameter estimation, reduced floor and ceiling effects as compared to non-adaptive procedures (Rock et al., 1995).

Another study, also led by Rock, is the Early Childhood Longitudinal Study (ECLS; Rock, 1996). In this study, Rock considered having a different test for each grade level or using a procedure similar to that utilized in the NELS:88 study – a delayed two-stage testing procedure that routed the child to a form based on how he or she performed on the previous year's testing. However, he realized that in the case of testing young children, these alternatives would not be optimal, as one could expect great variability among the children that is typically not reflected in a standard test. Instead, he used an individual semi-adaptive procedure that is commonly employed when testing young children. In such a procedure, each child is given approximately five items targeted at his or her ability level and, depending on the child's performance on this first stage, the child is given an ordered sequence of more difficult items or an easier sequence of items until he or she reaches a set number of consecutively incorrect responses. A drawback of this procedure is the lack of standardization in terms of the reliability and content coverage of each test due to the differing number of items each child will see.

The last example of a multi-stage test comes from the National Board of Medical Examiners (NBME). Their MST program is Computer Adaptive Sequential Testing (CAST). Implementation of CAST in the United States Medical Licensure Examination (USMLE) is expected to begin with the April/May 1999 administration of USMLE's Step 1 exam, followed by the Step 2 Exam in July/August 1999, and the Step 3 Exam sometime later in 1999 (Federation of State Medical Boards of the United States, Inc. & National

Board of Medical Examiners, 1998). The USMLE is a three-step exam, which is designed to assess a physicians' ability "to apply knowledge, concepts, and principles that are important in health and disease and that constitute the basis of safe and effective patient care" (Federation of State Medical Boards of the United States, Inc. & National Board of Medical Examiners, 1998, p. ii).

"CAST is a structured approach to test construction which incorporates adaptive testing methods with automated test assembly to allow test developers to maintain a greater degree of control over the production, quality assurance, and administration of different types of computerized tests" (Luecht & Nungester, 1998, p. 2). It appears that the NBME made a conscious decision to engage in MST rather than CAT (Luecht et al., 1996; Luecht & Nungester, 1998). Their primary reason was that with MST, it is more feasible to have test specialists and committees review test forms as there are usually a small fixed number of forms in MST as compared to the large number of test forms possible in CAT. In high stakes medical testing, having humans review test forms is considered highly necessary by medical test policy-makers and researchers.

From these examples it is clear that there are advantages associated with MST. The advantages, as well as the disadvantages are described below, respectively.

2.3.2 Advantages of MST

A distinct advantage of MST is the choice of delivery mode of the test. Unlike computerized adaptive tests that are assembled in real-time while the examinee is taking the test (and thus need to be administered by computer), multi-stage tests are assembled well before administration of the test and thus, may be delivered via computer or by P&P.

When administered by P&P, first-stage scoring can be accomplished by simply using total right scoring or by attaching a weight to each item based on its difficulty and then calculating total score. Alternatively, if stage two is delayed, as in the NELS:88 (Rock et al., 1995) and ECLS (Rock, 1996), IRT scoring can be used.

When a multi-stage test is given by P&P, it has the additional advantage of administering the test in a group rather than individually. In theory, one could also deliver a computerized multi-stage test in a group. However, in practice, the number of computers and electrical outlets that would be needed for a group administration would exceed the capacity of any testing center or gymnasium. The advantages of a group administration of a test are that it is more efficient and economical in terms of the number of people you can test in a set amount of time, and that the item pool would need not be as large. It is well known that a testing program which offers the convenience of year-round testing requires a large item pool for reasons of test security to keep item exposure to a minimum. While there are additional advantages to be gained by administering a multi-stage test via P&P in a group, one would no longer be able to capitalize on any of the advantages of computerized testing.

When a multi-stage test is administered by computer, it maintains all of the advantages of CAT – increased accuracy in measurement through the computer's capacity to support new and innovative item formats and the potential to improve test security, providing additional information for assessing proficiency from the speed of response, increasing economy of paper, improved data collection and pretesting of items, year-round testing, the convenience and flexibility of individual scheduling, immediate feedback to examinees (which can be beneficial for diagnostic purposes), and expedited score

reporting services (Wainer, Dorans, Flaugher, Green, Mislevy, Steinberg, & Thissen, 1990).

Whether MST is accomplished via computer or via P&P, it has the advantages derived from the adaptive nature of the test, such as a potentially shorter test. Although MST does not entail the same degree of adaptation as CAT, it maintains the advantage of employing a potentially shorter test since items are still tailored to an examinee's ability level. Luecht, Nungester, and Hadadi (1996) found that a 600 to 630 item licensing examination could be reduced to a content balanced 180-item multi-stage test without a significant loss in precision. In fact, attenuation of reliability was only 0.03. Considering a 70% decrease in test length, they felt that moving from a reliability of 0.97 (the typical full length reliability) to 0.94 was justified and a desirable result. Furthermore, they found that the false positives obtained from a shorter multi-stage test closely approximated those found in a full-length test.

A second advantage derived from the adaptive nature of MST is heightened ability to better control item exposure and meet content specifications than CAT (Luecht & Nungester, 1998). Item exposure is better controlled due to the fact that only the existing forms of a multi-stage test – rather than the entire item pool – are at risk of being exposed on any particular day of MST at the testing sites; thus fewer items are exposed (Luecht, et al., 1996). With regard to content balancing, it is possible to content balance a multi-stage test while still maximizing reliability or minimizing decision and satisfying exposure controls (Luecht et al., 1996).

A third advantage derived from the assembly of tests “behind the scenes” is that only a fixed number of test forms are produced. This allows the opportunity to implement

quality assurance by having humans review test forms. However, one must view quality assurance as a continuum rather than as a procedure that one does or does not do. On one end of the continuum, humans have full control over quality assurance. This may be inefficient in terms of cost and time and could lead to inconsistencies over time. However, humans may also catch important errors like near duplicate items, items measuring duplicate concepts, or other fuzzy features that have not been or can not be coded or quantified and therefore can not be resolved by a computer algorithm.

On the other end of the continuum, the computer has full control over quality assurance in the sense that quality assurance is fully automated by computer. This is fast and efficient. Unfortunately, from a quality assurance perspective, the computer is only as good as the codes, rules, and algorithms incorporated into the item algorithm software. Computer automation is very good at meeting content and information constraints in the item selection process, hence, we have such programs as ConTEST to design optimal tests (van der Linden & Boekkooi-Timminga, 1989). However, it is not good on the quality assurance side (i.e., garbage in, garbage out).

There needs to be a compromise between the two methods of quality assurance. There are some things that humans do best (like resolving fuzzy logic or problems involving incomplete, ambiguous or inconsistent data) and some things that computers do best (like running high-speed computations and repetitive checking tasks on coded or numeric data). The key is to find a blend that guarantees top quality test forms over time, without high costs or wasted time.

MST allows one to implement quality assurance as they wish. A multi-stage testing program could function in a fully automated fashion, or they could have

committees review every test form. For example, the National Board of Medical Examiners will want committees to review most of the test forms until they become confident that the computer is doing the proper job 99.9% of the time (personal communication, R. Luecht, February 18, 1998). Subsequently, they may just have committees audit the occasional test or subtest.

In addition to the selection of delivery mode, the ability to better control item exposure, and the opportunity to implement quality assurance as desired, MST allows for ease of item review. Because examinees are administered fixed subtests, they can review and alter responses within a particular subtest.

Finally, a further advantage of MST is that it capitalizes on the use of existing automated test assembly procedures (Luecht et al., 1996; Luecht & Nungester, 1998).

In summary, the advantages of MST that go beyond those of CAT are allowing for choice of delivery mode, better control of item exposure and content balancing, the opportunity to implement quality assurance as desired, allowance for item review, and the capitalization on existing automated test assembly procedures.

2.3.3 Disadvantages of MST

One criticism of MST by testing programs that use CAT is that MST does not cover a broad ability spectrum and, thus, does not produce accurate ability estimates across the entire range of ability. However, Luecht, Nungester, and Hadadi (1996) have shown that MST is “nearly as good as CAT and could be made statistically better by simply manipulating the target information functions” (p. 18) of the subtests at each stage.

2.4 Comparison of CAT and MST

Computerized MST (CMST) is a special case of CAT, which falls between the two extremes of linear computer-based testing (LCBT) and CAT and is a compromise between the two. Among the three types of computerized testing (LCBT, CAT, and CMST), CAT and CMST are what many testing programs are striving to implement today. Since in both CAT and MST tests can be targeted to either maximize measurement precision or minimize pass/fail decision errors, it is a matter of weighing the advantages and disadvantages of each in deciding which to implement.

There are many studies that have compared CAT and CMST (Kim & Plake, 1993; Luecht et al. 1996; Schnipke & Reese, 1997). Each of these studies is described and discussed below.

The purpose of Kim and Plake's (1993) simulation study was twofold. Their primary purpose was to compare the accuracy and efficiency of ability estimates obtained from two-stage testing and CAT. Their secondary purpose was to determine under what conditions might two-stage tests yield ability estimates as accurate as those yielded by computerized adaptive tests. First, 1,600 examinees were generated by creating 100 thetas at each of 16 discrete ability levels at and between -3.0 and 3.0 . Subsequently, they constructed eighteen two-stage tests corresponding to three factors: length of routing test (10, 15, and 20), distribution of item difficulties in the routing test (rectangular or peaked), and the number of second-stage tests (6, 7, and 8). The second-stage tests consisted of 30 items each. They also constructed three fixed-length CATs corresponding to the total length of the two-stage tests: 40, 45, and 50 items. To make direct comparisons between the two-stage tests and computerized adaptive tests, they used item

pools similar to each other and equivalent in size (354 items), and they used the same underlying statistical model (a modified one-parameter model) and maximum likelihood estimation procedure.

Kim and Plake (1993) found that a fixed-length CAT was superior to the two-stage tests of equivalent length in terms of accuracy and efficiency of ability estimates. In terms of the conditions under which two-stage tests produced the most accurate ability estimates, they found that the statistical characteristics of the routing test had a major influence on the accuracy of ability estimation. The longest routing test with a rectangular distribution of item difficulties, and the one with an odd number of second-stage tests produced the most accurate ability estimates.

Unlike Kim and Plake (1993) who simulated their data, Luecht, Nungester, and Hadadi (1996) used real data. In fact, their rationale for conducting their study was that so much research involving practical implementations of CMST or CAT has been limited to simulations that do not necessarily reflect realistic testing conditions such that the research may or may not generalize across different testing programs. In particular, they were interested in content balancing and item pool exposure in ability estimation and mastery decisions.

Luecht and his colleagues (1996) described and compared five methods of test construction/item selection procedures applicable to developing and administering computerized adaptive tests. The first three methods were variations of a computerized adaptive test that differed in their item selection algorithm: content-constrained CAT (CCON), heuristically content-balanced CAT (CBAL), and randomesque adaptive mastery testing with heuristic content-balancing (RNDQ). For descriptions of each refer to

Luecht, Nungester, and Hadadi (1996, p. 22). The last two methods were variations on a multi-stage test. One multi-stage test was designed to maximize the accuracy of ability estimates for most examinees (ACAST). It consisted of four stages with 1, 2, 3, and 4 subtests per stage, respectively. The second multi-stage test was designed to minimize mastery decision errors (MCAST) and consisted of three stages with 1, 3, and 5 subtests per stage, respectively. To make direct comparisons between the multi-stage test and computerized adaptive tests, the same number of items ($n=180$), the same item pool, and the same scoring was used for each test (maximum likelihood estimation and expected a posteriori estimation). To make the simulation realistic, an item pool consisting of 2538 previously used items that were calibrated using the one-parameter model and placed on a common scale, 60 simultaneous content constraints, and empirical ability estimates from 20,000 examinees who were administered the items previously were used for the study.

Their results showed very little difference in mastery decisions among the five test construction methods and that all five methods produced ability estimates which closely approximated the true abilities. Because CCON represents the optimal statistical method of targeting items to examinees, they used it as the baseline for comparing the efficiency of the other methods. They found that CBAL was just as efficient, if not more efficient than CCON and that both ACAST and MCAST were about 90% as efficient as CCON throughout the middle ability range. This is acceptable given the additional advantages of MST over CAT. However, both ACAST and MCAST dropped at the extremes of ability ranges. RDNQ was the least efficient relative to CCON and showed the greatest decline in efficiency beyond the cut-point. Finally, in terms of item exposure rates, more items were exposed by the CAT methods than by the MST methods.

Finally, rather than using a computerized adaptive test as a baseline, Schnipke and Reese (1997) used two P&P tests as a baseline to compare the precision of ability estimates obtained from three multi-stage test designs to those derived from two computerized adaptive test designs. The multi-stage test designs consisted of a two-stage testlet design that rerouted examinees within the second stage as needed, and a four-stage testlet design. The two computerized adaptive test designs were a standard maximum-information item-level design (the psychometric ideal in terms of precision and efficiency) and a maximum-information testlet based design (adapted at the testlet level rather than the item level). One P&P test was the same length as the other designs (25 items) and the second P&P test was twice as long (51 items). To compare the seven test designs, a group of 25,000 examinees at each of 25 ability levels at and between -3.0 and 3.0 and item parameters based on a three-parameter model were simulated, and Bayes modal scoring was used.

Similar to Luecht, Nungester, and Hadadi (1996), Schnipke and Reese (1997) found that the item-level CAT design led to the least amount of error and bias in ability estimates, particularly in the tails of the ability distribution, than any of the other designs. The 25-item P&P design led to the greatest amount of error and bias in ability estimates. The two- and four-stage tests and the testlet-based CAT led to ability estimates that were quite similar in terms of error and bias to the 51-item P&P design for ability estimates less than 1.5. However, for ability estimates greater than 1.5, the 51-item P&P design outperformed the two- and four-stage designs.

2.4.1 Summary

It is clear that there are many advantages associated with both CAT and CMST, and that each is an improvement over P&P testing. Based on the studies that have compared CAT and MST (Kim & Plake, 1993; Luecht, Nungester, & Hadadi, 1996; Schnipke & Reese, 1997), it is apparent that a computerized adaptive test is more efficient than an equal length multi-stage test. CAT is more efficient in terms of reducing test length with no loss in measurement or decision accuracy. This is not surprising since items in CAT are selected item by item rather than by fixed sets of item within stages as in MST. However, statistical efficiency is not always the paramount concern in testing.

To those involved in high-stakes certification and licensure testing (e.g., medical licensure), there are additional criteria to be met that may take precedence over a purely statistical view of test construction (Luecht et al., 1996). Such criteria include tighter content balancing than is typically found in low stakes certification or licensure testing or non-certification or non-licensure testing and a greater need to control item exposure. These criteria are hard to meet simultaneously in CAT and therefore an alternative to CAT seems necessary. MST is one such alternative.

2.5 MST Design Factors

There are many factors that enter into the design of a multi-stage test. Lord (1980, p. 129) listed the following factors to consider in designing a two-stage testing procedure:

1. The total number of items given to a single examinee.
2. The number of alternative second-stage tests available for use.
3. The number of alternative responses per item.

4. The number of items in the routing test.
5. The difficulty level of the routing test.
6. The method of scoring the routing test.
7. The cutting points for deciding which second-stage test an examinee will take.
8. The difficulty levels of the second-stage tests.
9. The method of scoring the entire two-stage procedure.

With the exception of the number of alternative responses per item, all of these factors are still considered as important in designing a multi-stage test. If one would like to extend these factors to a more general multi-stage test, one need only replace second-stage test by stage test and consider the number of stages. In addition to these factors, one should also consider the ability distribution of examinees and the amount of overlap of stage tests at each stage. Each of these factors, and studies relating to each factor, are described herein.

1. Total number of items given to a single examinee. As with any test, increasing the number of items in a multi-stage test, by adding items of comparable quality to the existing items, will increase the reliability of the test and hence improve measurement by producing more accurate ability estimates (Loyd, 1984; Kim & Plake, 1993). However, one selling feature of a multi-stage test is that, in general, a multi-stage test requires fewer items than a P&P test to attain a comparable level of precision. In a study that examined balancing item information, content, and exposure; Luecht, Nungester, & Hadadi (1996) found that a 600-item P&P licensing exam could be reduced to a content-balanced 180-item multi-stage test with reliability dropping only marginally from 0.97 to 0.94 and the number of false positive errors remaining approximately the same.

A question concerning the optimal total number of items in a multi-stage test that does not appear to be addressed in the MST literature is: Given a test of fixed length, how many items should be included in each stage?

2. Number of items in the routing test. The number of items in a routing test can have a profound effect on the routing of an examinee to the second-stage test, and possibly to other stages as well, and can ultimately affect the accuracy of ability estimation. With the total number of items fixed, if the routing test is too long there will not be enough items for the second-stage test and in essence the test will no longer be adaptive. On the contrary, if the routing test is too short there will be poor allocation to a second-stage test and possibly poor measurement of ability (Lord, 1980). Both Loyd (1984) and Kim and Plake (1993) varied the number of items in the routing test, but neither kept total test length constant. Not surprisingly, a result of both studies was that the longer routing test was superior.

The purpose of Loyd's (1984) study was to compare the consistency with which three routing test lengths of 10, 15, and 20 items assigned examinees to the same second-stage test and the accuracy with which they indicated the most appropriate second-stage test. The routing test and one of six 40-item second-stage tests were administered to 1439 students. The 20-item routing test was most effective in routing examinees to appropriate second-stage tests.

Kim and Plake (1993) did not want to compare different routing tests, but instead wanted to determine under what conditions a two-stage test might be comparable to a computerized adaptive test in terms of accuracy of ability estimates. Among other factors, the two-stage tests simulated varied in terms of

length of routing tests (10, 15, or 20). They found that the longest routing test (20 items) produced the most accurate ability estimates. Neither Loyd (1984) nor Kim and Plake (1993) kept total test length constant. A question of interest is: With total number of items in a test fixed, what portion of items should be placed in the routing test?

3. Difficulty level of the routing test. There is consensus among practitioners that the routing test should be of moderate difficulty. However, another question of concern is what type of distribution should the difficulty of items in the routing test have? Moreover, what type of distribution should the information of items in the routing test have? Although Luecht, Nungester, and Hadadi (1996) considered target information functions in designing multi-stage tests, they did not examine the effect of varying information functions. Kim and Plake (1993) found that given a rectangular distribution of ability of examinees, a rectangular distribution of difficulties produced more accurate ability estimates than a peaked distribution of difficulties. Kim and Plake (1993) only examined the case of a rectangular distribution of examinees' ability. One would hypothesize that the distribution of difficulty of the routing test would be strongly influenced by the ability distribution of examinees. Given a peaked distribution of examinee ability, a routing test with a peaked distribution of item difficulty would be highly desirable. It is common that a testing organization knows the expected ability distribution of its target population. This information is an important factor in a multi-stage test design, but to date has not yet been investigated.

4. Deciding which stage test the examinee should take. It is important that the difficulty levels of the second-stage tests match the ability levels of the examinees allocated to them, as determined by the routing rules. Likewise, the difficulty levels of the third-stage tests should match the ability levels of the examinees allocated to them, as determined by the routing and second-stage test. In Lord's (1971) trial-and-error methods, he found that equally spaced cut-points on the routing test resulted in better allocation to a second-stage test than did unequally spaced cut-points. While others (e.g., Kim & Plake, 1993) seemed to arbitrarily set cut-points, Schnipke and Reese (1997) cleverly used the mean squared errors of ability estimates from each subtest to determine cut-points for routing examinees to various stage tests. One could also route an examinee to the stage test that provides the most information given the examinee's ability estimate from previous stages. More research is needed in this area.
5. Number of stages. While many researchers have studied the accuracy of ability estimates obtained from two- and four-stage tests (Lord, 1980; Loyd, 1984; Kim & Plake, 1993; Luecht et al., 1996; Schnipke & Reese, 1997), no researcher has examined the effects of systematically varying the number of stages on the accuracy of ability estimates or mastery decisions.
6. Number of stage tests per stage. Several studies have shown that the number of second-stage tests influences the quality of measurement (e.g., Lord, 1971b; Kim & Plake, 1993). Lord (1980) stated that there cannot usefully be more second-stage tests than the number of items in the routing test and that at least four subtests were required to achieve good measurement over the entire ability range.

Interestingly, Kim and Plake (1993) found that an odd number of second-stage tests produced the most accurate ability estimates. No one has investigated the effect of the number of stage tests at higher stages on ability estimation. More research is needed on the effects of the number of subtests per stage on the accuracy of ability estimation (Luecht et al., 1996; Luecht & Nungester, 1998).

7. Difficulty levels of the stage tests. Lord (1971b) found that if the difficulty levels of the second-stage tests are too close to the level of the difficulty level of the routing test, poor measurement is obtained at the extreme ability levels and if the difficulty levels are too extreme, there is poor measurement where the ability level of the examinee was too near the difficulty level of the routing test. In addition to the difficulty levels of the subtests, a concern is the distribution of difficulty levels of the stage tests on final ability estimates. No research has been done in this area.

One hypothesizes that the distribution of difficulty levels of the second-stage tests should probably be peaked according to the cut-points on the stage tests. Similarly, the distribution of difficulty levels of the third-stage tests should be peaked according to cut-points on the second-stage test. One may foresee the problem that can arise from peaking item difficulty of a stage test according to cut-points on the previous stage test and peaking it according to cut-points on subsequent stage tests. Such a method would lead to a multi-modal distribution of item difficulties, which would behave similarly to having a uniform distribution of item difficulties. Hence, a uniform distribution of item difficulties is desirable.

Alternatively, one may want to consider the distribution of information of the stage tests rather than distribution of item difficulties. In this way, information

from item discrimination and guessing is taken into account along with information from the item difficulty.

8. Overlap of difficulties of stage tests at the same stage. If, for example, the routing test is fairly long, then overlap of difficulties in stage tests at the next stage may not be very important. However, the overlap may be very important if the routing test is short since there will be less confidence in routing examinees from the first to second stage. There appear to be no studies in the literature that have investigated the effect of the amount of overlap on the accuracy of ability estimation or mastery decisions. Kim and Plake (1993) employed 60% of their items overlapping with adjacent subtests with the exception of the highest and lowest stage tests which only shared 30% of their items in common. Luecht, Nungester, and Hadadi did not simulate overlap in items; however, they did simulate overlap in item information. More research is needed to determine the optimal amount of overlap in order to minimize usage of the pool while maximizing the accuracy of ability estimates and mastery decisions.
9. Method of scoring the total test. In essence, there are n ability estimates for each examinee for a multi-stage test with n stages. These estimates are jointly sufficient statistics for the ability estimate of the total test and must be combined into a single estimate. Unfortunately, there is no unique, best way to obtain a final ability estimate (Lord, 1980). Lord (1980) suggests averaging all of the ability estimates after weighting them inversely according to their estimated large sample variances as it "is well known that this weighting produces a consistent estimator with approximately minimum large-sample variance" (Lord, 1980, p. 131). Today, it is

most common to combine all items in a multi-stage test and estimate ability with Lord's method. In doing this, Luecht, Nungester, and Hadadi (1996) found very little difference between expected a posteriori (EAP) and maximum likelihood estimation (MLE) of ability. Most recent studies in the MST literature utilized MLE scoring procedures (Loyd, 1984; Kim & Plake, 1993; Schnipke & Reese, 1997).

2.6 Purpose of Study

While there are many variables to consider when designing a multi-stage test, in this study some variables were fixed to examine the effects of varying other variables on the accuracy of ability estimates produced by MST. The ability estimates obtained from various designs of MST were compared with those obtained from CAT and a P&P test. Total test length was kept constant. While Kim and Plake (1993) studied the number of second stage tests systematically, they also varied test length by varying the length of the routing test. In addition, although Luecht, Nungester, and Hadadi (1996) and Schnipke and Reese (1997) investigated MST designs with different numbers of stages and different numbers of subtests per stage with test length kept constant, neither factor was studied systematically. Furthermore, no study in the literature investigated the effect of the number of items in each stage test. The number of stages, the number of subtests per stage, and the number of items in each subtest were manipulated in this study. Finally, while previous studies examined the effect of item difficulty distribution in the stage tests,

this study, similar to the Luecht, Nungester, and Hadadi (1996) study, examined the effect of the distribution of item information in stage tests.

A simulation study, using item parameters from a real item pool and ability parameters based on three-parameter logistic calibrations of real data, was conducted in which total test length and the amount of overlap between stage tests was kept constant. In addition, item information was considered rather than item difficulty. Finally, based on the finding that there was little difference in methods of scoring the total test (Luecht et al., 1996), the method of scoring the total test was held constant. The following factors were varied:

- a. number of stages,
- b. number of stage tests per stage, and
- c. number of items per stage test and routing test.

The primary question of interest was, given a fixed test length, how many stages and how many subtests per stage should there be in order to maximize measurement precision? Furthermore, given a fixed test length, how many items should there be in each subtest? Should there be more in the routing test? Or should there be more in the higher stage tests? A secondary question of interest concerned conditional item exposure rates and the number of items exposed by CAT and the different MST designs.

CHAPTER 3

METHODOLOGY

In this chapter, the methodology for the study is presented. The method is divided into four sections: test conditions, computer programs, procedure, and data analysis.

3.1 Test Conditions

Data corresponding to 14 different test conditions were simulated. Each test condition corresponded to one test design: a P&P test (1), a computer adaptive test (1), or a multi-stage test (12). To allow for a direct comparison between test designs in this study, all test designs had a fixed test length that were selected from the same precalibrated (using IRT's 3PL model) item pool.

In order to avoid confounding the study with item parameter estimation issues and to make the simulations realistic, an existing item pool consisting of 1256 precalibrated items from the Logical Reasoning section of the Law School Admission Test (LSAT) was used. All items in the pool were multiple-choice items with five alternative responses and were coded with respect to their content. To reflect what is commonly found in practice, a 36-item test was used and the 1256-item pool was partitioned into three parallel subpools of 418 items each and only one subpool was used in this study.

Furthermore, so that the test conditions would reflect what occurs operationally, each of the tests was designed to satisfy certain item exposure constraints and content specifications. Item exposure was controlled conditionally at 10 ability levels in CAT and for building the multi-stage tests. Conditional exposure rates of .20 to .38 are commonly

found among operational CAT programs (e.g., GMAT and GRE). In this study, the maximum conditional item exposure rate was set to .25. In addition, the same nine LSAT content constraints were used in all test designs. They are outlined in Table 3.1.

Table 3.1
Content Specifications for all Test Designs

Content Category	Number of Items
1	2
2	5
3	4
4	1
5	5
6	1
7	7
8	8
9	3
Total	36

Finally, a maximum likelihood estimation procedure was used. Each test was administered to 500 simulated examinees at each of 10 ability levels (-2.25 to 2.25 in increments of .5). In this way, the 500 replications at each ability level provided a convenient and accurate basis for estimating measurement precision. Estimation procedure was not varied in this study since Luecht, Nungester, and Hadadi (1996) found relatively little difference between expected a posterior (Bayes) and maximum likelihood ability estimates.

There was one P&P design, one CAT design and 12 MST designs. The P&P design would result in the poorest results, and the CAT design would produce the best

results. The P&P and CAT designs were used as the bases for comparing the MST designs. Each design is described in more detail, next.

3.1.1 P&P Condition

To allow for a direct comparison among the P&P, CAT, and the MST designs, the same content constraints and 418-item subpool were used to construct a 36-item P&P test. Based on advice from LSAT test specialists who stated that they like to build 12 51-item tests from their 1256-item pool (S. Luebke, personal communication, October, 28, 1998), the target in this study was to build five 36-item tests using the 418-item subpool. To avoid building a single “best” P&P test for use in this study, information from a typical CAT, with maximum conditional item exposure set to .20, was used to specify a target information function for the test.

3.1.2 CAT Condition

Items for the 36-item computerized adaptive test were selected from the 418-item subpool for each examinee based on the maximum information item selection procedure, as specified by the 3PL IRT model and subject to the same content constraints as were used with the P&P and MST designs and a maximum conditional exposure rate of .25. To ensure that each test satisfied the nine content constraints, once enough items from a content category were delivered to an examinee, items from that category were “shut off” and were no longer available to be selected for that examinee.

To avoid all of the “good” (highly discriminating, highly informative) items from being administered to the first 100 examinees, a 2:1 weighting was given to information

and exposure, respectively, in the item selection procedure. A selection value for each item available for selection was calculated based upon the examinee's ability estimate and the number of times the item had been exposed. The selection value for item i in the subpool, given examinee j with ability θ , was the weighted sum of information for item i given examinee j with ability θ and an indicator of exposure, xp_i :

$$Selection_i = 2 * I_i(\theta_j) + 1 * xp_i, \quad (3.1)$$

where

$$xp_i = \sqrt{\frac{xp_{max} - \text{exp}}{xp_{max} - xp_{min}}} \quad (3.2)$$

and

exp = exposure rate,

xp_{max} = maximum exposure rate desired

xp_{min} = minimum exposure rate desired

Item information, $I_i(\theta_j)$ was calculated at 60 θ values (examinee ability estimate +/- 3.0 in increments of 0.1) for each available item i using the formula

$$I_i(\theta_j) = \frac{2.89a_i^2(1 - c_i)}{\left[c_i + e^{1.7a_i(\theta_j - b_i)} \right] \left[1 + e^{1.7a_i(\theta_j - b_i)} \right]^2}, \quad (3.3)$$

where a_i , b_i , and c_i are discrimination, difficulty, and guessing parameters for item i ,

respectively and θ_j is the j th ability value (Hambleton, Swaminathan, & Rogers, 1991).

The item with the greatest selection value corresponding to the examinees's estimated θ was selected and administered.

With a 2:1 weighting of information and exposure, respectively, a less informative item with a smaller exposure rate could be selected for administration in lieu of a more informative item with an exposure rate close to 1.0. For example, with maximum and minimum exposure set to 0.25 and 0, respectively, Item 2 in Table 3.2 would be selected over Item 1, regardless of Item 1 being more informative than Item 2.

Table 3.2

Example of Item Selection

Item	Information	Exposure	xp	Selection
1	.80	.20	.45	2.0
2	.75	.10	.60	2.1

Finally, the examinee's θ estimate was updated after each item was administered using a maximum likelihood estimation procedure. The examinee's θ estimate after 36 items were administered was used as the examinee's final θ estimate.

3.1.3 MST Conditions

To allow for a direct comparison between MST designs and the P&P and CAT designs, multi-stage panels were constructed using the same calibrated 418-item subpool, taking into account the same content constraints as were used in the P&P and CAT designs. A panel is the combination of subtests for a particular MST design. To avoid building the "best" panel given the 418-item item pool, for each MST design, information from a typical CAT with maximum conditional item exposure set at .25 was used to build target information functions for each subtest. In addition, for each MST design, two multi-

stage panels were built. Each MST design was defined by a combination of three factors: (1) number of stages, (2) number of subtests per stage, and (3) number of items per subtest.

3.1.3.1 Number of Stages

Two- and three-stage tests were constructed for this study. Two is the minimum number of stages to have in a multi-stage test and is what is most commonly found in practice and in the literature (Kim & Plake, 1993; Loyd, 1984; Luecht & Nungester, 1998; Rock, 1996; Rock, Pollack & Quinn, 1995; Schnipke & Reese, 1997). Three stages was chosen as the upper limit as the subpool could not support the number of items required for more stages.

3.1.3.2 Number of Subtests per Stage

Three and five subtests for each of the second and third stages were constructed. The lower limit of three subtests per stage was chosen because it is what is commonly found in practice and aside from pass/fail decisions, which might require only two subtests, seems to be the minimum number of subtests one would desire. Five subtests per stage was chosen as the upper limit as it seems to be the maximum number of subtests per stage to which examinees could meaningfully be assigned and the maximum number which the subpool could support.

3.1.3.3 Number of Items per Subtest

The number and percent of items per subtest per stage varied depending on the number of stages in the multi-stage test design. (Note that all subtests in a given stage contained the same number of items.) In this study, the number of items per subtest per stage were varied based on three rationales. A summary of these numbers can be found in Tables 3.3.1 and 3.3.2.

One rationale for allocating the number of items to a stage is, assuming that there is more precise measurement in the higher stages (due to examinees receiving more items that are more closely matched to their ability in higher stages), more items should be placed in the higher stages. Based on this rationale, the percent of items in the two-, three-stage tests, respectively, were $1/3$ and $2/3$ and $1/6$, $1/3$, and $1/2$, respectively. This rationale is referred to as the “Higher Stage” rationale in Tables 3.3.1 and 3.3.2.

A second rationale that is in opposition to the first rationale is that there is a need for accurate measurement in the routing test in order to properly allocate examinees to subsequent stages and therefore, more items should be placed in the routing test than in the other stage tests. Based on this rationale, the percent of items in the two- and three-stage tests, respectively, were $2/3$ and $1/3$ and $1/2$, $1/3$, and $1/6$, respectively. This rationale is referred to as the “Routing Test” rationale in Tables 3.3.1 and 3.3.2.

Finally, as a compromise of the above two rationales and for completeness purposes, the effect of placing equal numbers of items in each stage was also investigated. Based on this rationale, the percent of items in a two- and three-stage test, respectively, were $1/2$ and $1/2$ and $1/3$, $1/3$, and $1/3$, respectively. This final rationale is referred to as the “Compromise” rationale in Tables 3.3 and 3.4.

Table 3.3

Proportion of Items Per Subtest Per Stage

Rationale	Number of Stages				
	Two		Three		
	1	2	1	2	3
Higher Stage	1/3	2/3	1/6	1/3	1/2
Compromise	1/2	1/2	1/3	1/3	1/3
Routing Test	2/3	1/3	1/2	1/3	1/6

Table 3.4

Number of Items Per Subtest Per Stage

Rationale	Number of Stages				
	Two		Three		
	1	2	1	2	3
Higher Stage	12	24	6	12	18
Compromise	18	18	12	12	12
Routing Test	24	12	18	12	6

As seen in Tables 3.3 and 3.4, the interaction of the number of stages and the number of items per subtest per stage resulted in six test conditions.

3.1.3.4 Summary

In summary, the interaction of the number of stages (2 or 3 stages) and the number of items per subtest (3 rationales), and the two levels of subtests (3 or 5 subtests per stage) yielded 12 (2x3x2) MST designs (see Table 3.5).

Table 3.5

Summary of MST Designs

MST Design	Number of Stages	Number of Subtests			Number of Items		
		Stage1	Stage2	Stage3	Stage 1	Stage 2	Stage3
I	2	1	3	--	18	18	--
II	2	1	3	--	12	24	--
III	2	1	3	--	24	12	--
IV	3	1	3	3	12	12	12
V	3	1	3	3	6	12	18
VI	3	1	3	3	18	12	6
VII	2	1	5	--	18	18	--
VIII	2	1	5	--	12	24	--
IX	2	1	5	--	24	12	--
X	3	1	5	5	12	12	12
XI	3	1	5	5	6	12	18
XII	3	1	5	5	18	12	6

3.2 Computer Programs

In this section, the computer programs that were used to construct the tests and simulate examinees through the different test designs are described. CASTISEL (Luecht, 1996) was used to construct the P&P and MST designs and LCMS90 (Robin & Patsula, 1998) was used to simulate examinees.

A modified 3PL version of CASTISEL (Luecht, 1996) was used to construct the MST tests. CASTISEL works with a calibrated item pool and allows the user to specify the total test length, the number of stages, the number of subtests per stage, and the number of items per subtest to construct a test. The program uses a local optimization heuristic, the “normalized weighted absolute deviation” (Luecht & Hirsch, 1992; Luecht, to appear) to fit the items to the information functions specified by the user, subject to precisely meeting the content constraints. Additionally, CASTISEL uses a matrix

partitioning algorithm (Luecht & Nungester, 1996) to optimally assign content constraints to the stages, so that content is balanced at each stage, as well as at the total test level. In this study, CASTISEL was used to construct five forms of a 36-item P&P test and two panels each of the 12 multi-stage test designs.

The P&P, CAT, and MST simulations were performed using LCMS90 (Robin & Patsula, 1998). LCMS90 allows the user to specify a fixed length or variable length CAT with an appropriate stopping rule, a linear-on-the-fly test, or a multi-stage test. It also allows the user to read or generate item and ability parameters with varying distributions and to input content specifications and exposure controls. Finally, it allows the user choose between methods of scoring the total test. In this study, LCMS90 was used to simulate 500 examinees at each of 10 ability levels through 36-item P&P, computer adaptive, and multi-stage tests. Item parameters were read from a precalibrated item pool and each test was designed to meet certain content and exposure constraints. Finally, a maximum likelihood estimation procedure was used to estimate examinee ability.

3.3 Procedure

In this section, the steps that were used to simulate the data and measure and compare the accuracy of ability estimation from the MST designs to those obtained from the P&P and CAT designs are described in detail.

3.3.1 Step 1 – Partition Item Pool

The 1256-item pool was partitioned into three parallel subpools of 418 items each by first sorting the items by content and then by difficulty and secondly, assigning the

sorted items to subpools one to three. This resulted in three content- and difficulty-equivalent subpools. As seen in Table 3.6, the mean and standard deviation of the item parameters were very similar across subpools. Furthermore, subpool information was comparable across all four subpools (see Figure 3.1). One subpool was randomly chosen and used for this study (Subpool 3).

Table 3.6

Descriptive Statistics of Item Parameters in the Total Pool and Each Subpool

Subpool	N	a		b		c	
		M	SD	M	SD	M	SD
1	419	.74	.24	-.09	1.14	.16	.10
2	419	.75	.25	-.05	1.14	.17	.10
3	418	.77	.25	-.07	1.13	.18	.11
Total	1256	.75	.25	-.07	1.13	.17	.11

3.3.2 Step 2 – Construct Tests

Whereas the computer adaptive tests were built in real time, on-the-fly by the computer, the P&P and MST tests needed to be built by hand. This required specifying target information functions for each subtest for each test. To avoid building the single best panel with maximum test information relative to the entire pool, the goal was to develop test information targets that reflected average or slightly less than average item pool information which could be maintained across multiple panels for a 36-item multi-stage test. This consideration becomes important when reviewing subsequent results. That is to say, if desired, it would have been relatively easy to design information targets for each subtest to essentially match the precision of a maximum-information CAT.

However, in order to be able to create multiple forms from the same pool, target test information functions were less than the maximum.

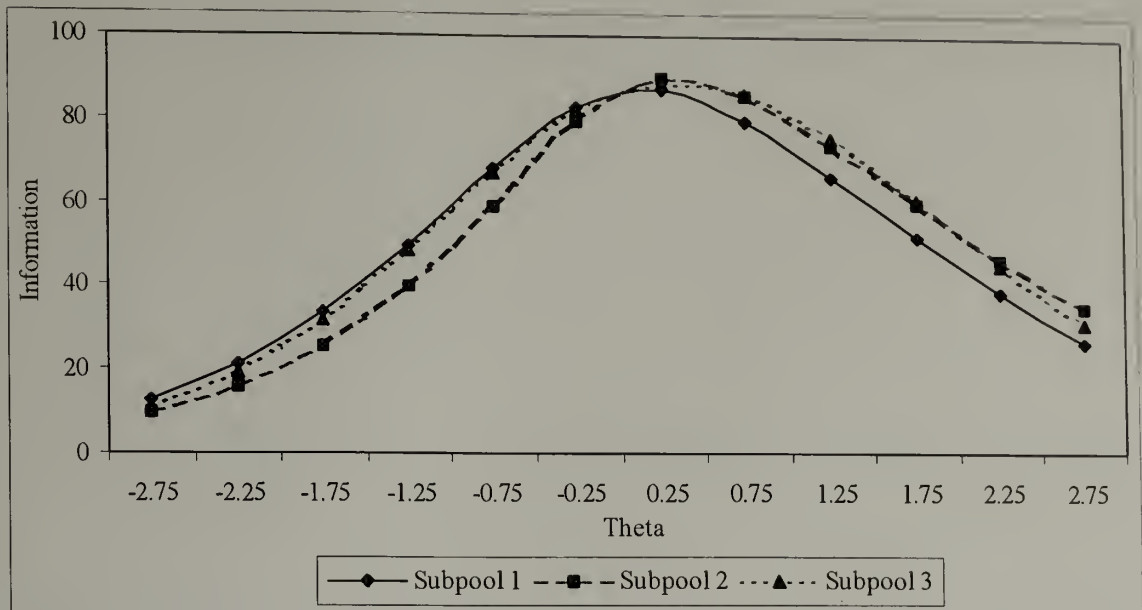


Figure 3.1. Subpool Information

In this study, two panels were created for each MST design. So as not to over or under specify target information functions for each subtest relative to the item pool, test information functions from a typical CAT with conditional exposure controls built in (.25) were used to build the subtest target information functions. Because test information functions are additive, the total test information function from a typical CAT could be divided into separate subtest information functions. To describe how the target information function was specified for each subtest, a two-stage test with three subtests in the second stage and 18 items in each subtest is used as an example (see Figure 3.2).

- i) For a two-stage test with three levels at the second stage, there are three pathways: an easy pathway (A+B), an average difficulty pathway (A+C), and a hard pathway (A+D).

- ii) Given these three pathways, the ability range was divided into three regions with each region consisting of an equal number of people. The midpoint of each region was then calculated. Using the LSAC ability distribution corresponding to the item pool, this corresponded to the following mid-points: -.72, .15, and .98.

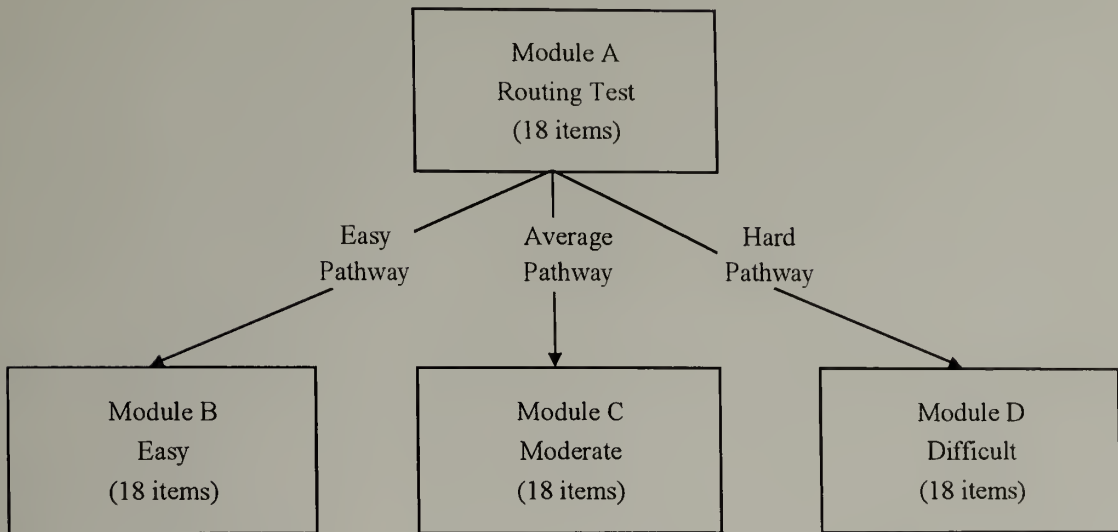


Figure 3.2. 36-Item Two-Stage Test with Three Subtests at the Second Stage

- iii) For Content Area One, 100 examinees at each of the region midpoints were administered a computerized adaptive test containing only items from Content Area One and the average test information function (TIF) was calculated for Content Area One for each region/pathway (TIF_1 , TIF_2 , and TIF_3).
- iv) To obtain the TIF for stage one (TIF_A), TIF_1 , TIF_2 , and TIF_3 were averaged together and multiplied by the proportion of items in stage one.
- v) To obtain TIF's for Subtests B, C, and D in stage two, TIF_A was subtracted from each average TIF pathway:

$$TIF_B = TIF_1 - TIF_A$$

$$TIF_C = TIF_2 - TIF_A$$

$$TIF_D = TIF_3 - TIF_A$$

- vi) Steps 3-5 were repeated for Content Areas 2 to 9.
- vii) Finally, the TIF's for each subtest were added across content areas (see Figure 3.3).

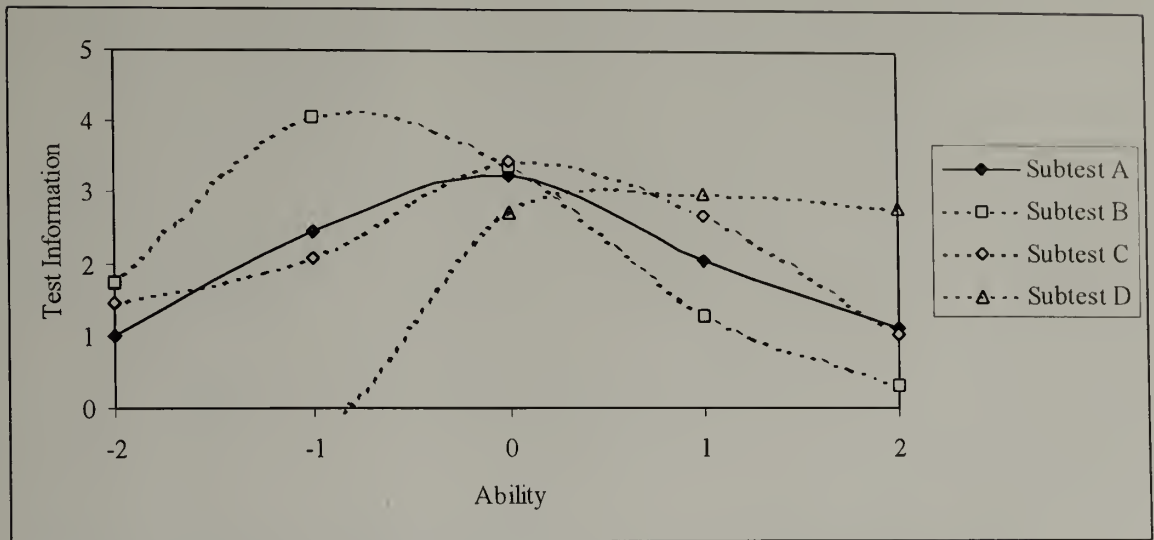


Figure 3.3. Example of a 36-Item Multi-Stage Test with 18 Items at Each Stage

For a multi-stage test with five subtests per stage, the ability range was divided into five regions. For three-stage tests, the TIF's for the subtests in stages two and three were weighted according to the proportion of items in each stage. After TIF's were specified for each subtest, CASTISEL was used to assemble the multi-stage tests.

Table 3.7 presents the average means and standard deviations of the 3PL item parameters for each subtest in the five forms of the P&P test and the two panels of each MST test design. The size of the standard deviations of the b 's relate to the relative shape of the observed TIF. That is to say, a larger standard deviation corresponds to a flatter curve whereas a smaller standard deviation corresponds to a more peaked curve which

reflects a more restrictive targeting of items to a particular region of the score scale. Interestingly, the standard deviations of the b 's are smallest for the middle difficulty subtests in Stage 2 and 3. This is most likely due to the fact that most items in the subpool are of middle difficulty (see Figure 3.4). To meet the “easy” (Subtest B) and “hard” (Subtest D) target TIF's, the assembly procedures selected easy and hard items, respectively, but also had to choose items of varying difficulty due to the restriction of choice of items. The mean of b 's indicate where each subtest provides maximum information.

Table 3.7 also summarizes the difference between the observed subtest information function of the selected items and the target for each subtest in the P&P and MST designs. The mean TIF difference is a simple average of the deviations for the number of items in a subtest computed across a grid of 31 points for θ ranging between -2.0 and 2.0 . The MSE (mean square error) of the TIF difference is the average squared deviation. With the exception of Subtest F in MST designs IV, V, and VI, the results indicate that the process of selecting items for each subtest to fit the target TIF's was fairly accurate. This was not unexpected as the item bank was specifically designed to provide maximum information for high ability examinees (see Figure 3.4).

3.3.3 Step 3 – Simulate Examinees

To generate item responses for each test design, a group of 5,000 examinees were simulated with 500 θ 's at each of 10 ability levels (-2.25 to 2.25 in increments of $.50$). These θ 's were treated as the true θ 's in the remainder of the study.

3.3.4. Step 4 – Simulate Item Responses

Finally, by utilizing LCMS90, each of the 14 test designs were administered to the 5,000 simulated examinees. The result was 5,000 θ estimates for each design, 500 independent ability estimates at 10 ability levels.

3.4 Data Analysis

According to the purpose of the study, the data were analyzed in two parts corresponding to the precision of ability estimation and item exposure. First, the analyses used to examine the precision of ability estimation are described. This is followed by a description of analyses used to investigate item exposure rates.

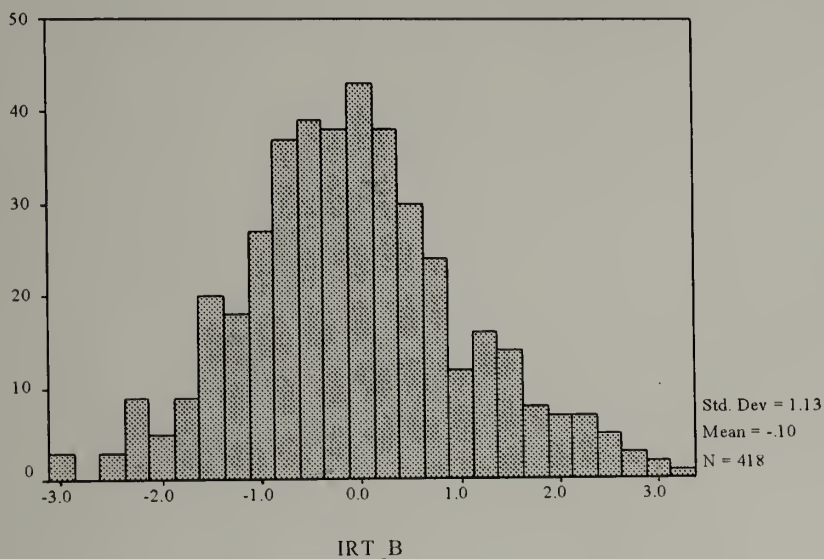


Figure 3.4. Frequency Distribution of b Parameters in the Pool

Table 3.7

CASTISEL Assembly Results for P&P and MST Designs

Design	Subtest	Stage	No. Items	Mean a	SD a	Mean b	SD b	Mean c	SD c	Mean TIF Difference	MSE of TIF Difference
P&P	1	1	36	0.55	0.07	-0.31	0.57	0.14	0.08	-0.04	0.10
P&P	2	1	36	0.59	0.16	-0.39	0.81	0.15	0.11	-0.05	0.11
P&P	3	1	36	0.61	0.15	-0.20	1.06	0.18	0.13	-0.04	0.09
P&P	4	1	36	0.64	0.16	-0.42	1.23	0.18	0.12	-0.08	0.19
P&P	5	1	36	0.71	0.23	-0.24	1.49	0.21	0.14	-0.09	0.16
I	1	1	18	0.60	0.13	-0.34	0.30	0.17	0.12	-0.03	0.01
I	2	2	18	0.72	0.16	-1.02	0.31	0.20	0.16	-0.01	0.07
I	3	2	18	0.64	0.15	-0.32	0.79	0.16	0.14	-0.04	0.03
I	4	2	18	0.68	0.18	0.60	0.85	0.18	0.10	-0.16	0.27
I	5	1	18	0.59	0.11	-0.33	0.59	0.17	0.10	-0.01	0.01
I	6	2	18	0.67	0.17	-0.94	0.68	0.12	0.07	-0.07	0.12
I	7	2	18	0.64	0.23	-0.37	0.72	0.16	0.14	-0.01	0.01
I	8	2	18	0.71	0.14	0.73	1.07	0.19	0.12	-0.18	0.31
II	1	1	12	0.60	0.05	-0.29	0.32	0.18	0.08	-0.02	0.01
II	2	2	24	0.67	0.15	-0.86	0.40	0.17	0.14	-0.06	0.03
II	3	2	24	0.63	0.15	-0.33	0.79	0.17	0.12	-0.02	0.01
II	4	2	24	0.62	0.19	0.24	0.51	0.17	0.11	-0.09	0.10
II	5	1	12	0.61	0.17	-0.41	0.40	0.16	0.13	-0.04	0.01
II	6	2	24	0.67	0.18	-1.05	0.57	0.15	0.10	-0.08	0.06
II	7	2	24	0.64	0.19	-0.48	1.08	0.16	0.12	-0.02	0.02
II	8	2	24	0.66	0.18	0.28	1.01	0.21	0.14	-0.10	0.10

Continued, next page.

Table 3.7, cont'd:

Design	Subtest	Stage	No. Items	Mean a	SD a	Mean b	SD b	Mean c	SD c	Mean TIF Difference	MSE of TIF Difference
III	1	1	24	0.61	0.12	-0.35	0.36	0.18	0.12	-0.02	0.01
III	2	2	12	0.79	0.14	-1.01	0.31	0.21	0.11	-0.10	0.03
III	3	2	12	0.66	0.12	-0.24	0.63	0.15	0.14	-0.05	0.04
III	4	2	12	0.60	0.11	0.77	0.67	0.14	0.10	-0.10	0.06
III	5	1	24	0.62	0.15	-0.40	0.53	0.18	0.13	-0.02	0.01
III	6	2	12	0.74	0.20	-1.09	0.46	0.17	0.14	-0.09	0.04
III	7	2	12	0.67	0.14	-0.32	0.83	0.15	0.11	-0.06	0.04
III	8	2	12	0.65	0.16	0.77	0.97	0.21	0.13	-0.07	0.07
IV	1	1	12	0.63	0.15	-0.40	0.29	0.21	0.13	0.00	0.00
IV	2	2	12	0.66	0.10	-0.90	0.34	0.18	0.09	-0.04	0.01
IV	3	2	12	0.64	0.11	-0.09	0.48	0.18	0.12	-0.03	0.02
IV	4	2	12	0.55	0.11	-0.07	0.69	0.10	0.08	-0.13	0.09
IV	5	3	12	0.65	0.17	-0.83	0.68	0.15	0.08	-0.02	0.01
IV	6	3	12	0.69	0.16	-0.53	0.91	0.15	0.12	-0.07	0.04
IV	7	3	12	0.73	0.19	0.63	1.01	0.23	0.13	0.02	0.02
IV	8	1	12	0.62	0.13	-0.41	0.45	0.19	0.13	-0.03	0.01
IV	9	2	12	0.71	0.18	-0.98	0.55	0.18	0.16	-0.08	0.03
IV	10	2	12	0.66	0.20	-0.07	0.61	0.18	0.15	-0.04	0.01
IV	11	2	12	0.56	0.16	-0.14	0.89	0.13	0.12	-0.08	0.06
IV	12	3	12	0.71	0.18	-0.95	1.01	0.19	0.14	-0.05	0.01
IV	13	3	12	0.69	0.21	-0.36	0.97	0.15	0.10	-0.06	0.04
IV	14	3	12	0.72	0.15	0.60	1.02	0.21	0.14	-0.03	0.04

Continued, next page.

Table 3.7, cont'd:

Design	Subtest	Stage	No. Items	Mean a	SD a	Mean b	SD b	Mean c	SD c	Mean TIF Difference	MSE of TIF Difference
V	1	1	6	0.59	0.08	-0.33	0.27	0.18	0.10	0.00	0.00
V	2	2	12	0.64	0.15	-0.79	0.26	0.18	0.13	-0.02	0.01
V	3	2	12	0.61	0.12	-0.21	0.45	0.15	0.11	-0.05	0.01
V	4	2	12	0.58	0.14	0.09	0.51	0.14	0.12	-0.08	0.04
V	5	3	18	0.64	0.18	-0.85	0.59	0.15	0.13	-0.06	0.03
V	6	3	18	0.60	0.19	-0.47	0.99	0.15	0.10	-0.01	0.01
V	7	3	18	0.68	0.20	0.25	0.89	0.22	0.15	-0.09	0.07
V	8	1	6	0.57	0.05	-0.38	0.33	0.15	0.06	-0.03	0.00
V	9	2	12	0.67	0.17	-0.95	0.50	0.18	0.16	-0.05	0.02
V	10	2	12	0.60	0.08	-0.40	0.76	0.15	0.06	-0.02	0.01
V	11	2	12	0.63	0.21	-0.14	0.93	0.17	0.14	-0.08	0.05
V	12	3	18	0.71	0.17	-0.94	0.97	0.16	0.12	-0.13	0.11
V	13	3	18	0.70	0.20	-0.45	1.27	0.20	0.13	-0.05	0.02
V	14	3	18	0.71	0.24	0.56	1.19	0.24	0.12	-0.06	0.03
VI	1	1	18	0.62	0.13	-0.39	0.32	0.19	0.12	-0.02	0.01
VI	2	2	12	0.68	0.14	-0.97	0.33	0.17	0.14	-0.07	0.02
VI	3	2	12	0.64	0.13	-0.20	0.60	0.18	0.12	-0.01	0.02
VI	4	2	12	0.57	0.16	0.45	0.30	0.13	0.09	-0.09	0.05
VI	5	3	6	0.67	0.28	-1.03	0.41	0.15	0.13	-0.03	0.00
VI	6	3	6	0.70	0.17	-0.54	0.94	0.22	0.12	0.00	0.00
VI	7	3	6	0.67	0.16	0.62	1.14	0.28	0.11	0.00	0.01
VI	8	1	18	0.60	0.12	-0.40	0.59	0.15	0.10	-0.05	0.01
VI	9	2	12	0.69	0.15	-1.11	0.53	0.17	0.13	-0.08	0.03
VI	10	2	12	0.69	0.21	-0.31	0.81	0.19	0.16	-0.02	0.01
VI	11	2	12	0.62	0.20	0.43	0.78	0.19	0.13	-0.06	0.04
VI	12	3	6	0.73	0.16	-1.17	0.81	0.14	0.10	-0.15	0.05
VI	13	3	6	0.58	0.24	-0.26	0.60	0.13	0.11	0.00	0.01
VI	14	3	6	0.61	0.19	0.37	0.84	0.16	0.08	-0.08	0.03

Continued, next page.

Table 3.7, cont'd:

Design	Subtest	Stage	No. Items	Mean a	SD a	Mean b	SD b	Mean c	SD c	Mean TIF Difference	MSE of TIF Difference
VII	1	1	18	0.63	0.08	-0.21	0.30	0.14	0.08	0.02	0.03
VII	2	2	18	0.66	0.13	-0.63	0.66	0.11	0.09	-0.14	0.37
VII	3	2	18	0.71	0.12	-0.25	0.54	0.16	0.09	0.01	0.03
VII	4	2	18	0.85	0.21	-0.27	0.59	0.25	0.16	-0.04	0.11
VII	5	2	18	0.80	0.19	0.03	0.57	0.24	0.12	0.07	0.24
VII	6	2	18	0.76	0.23	1.12	0.88	0.19	0.10	0.12	0.10
VII	7	1	18	0.64	0.07	-0.21	0.38	0.14	0.09	0.01	0.04
VII	8	2	18	0.69	0.19	-0.76	0.77	0.14	0.09	-0.13	0.38
VII	9	2	18	0.71	0.19	-0.29	0.63	0.15	0.12	0.00	0.03
VII	10	2	18	0.75	0.25	-0.42	0.84	0.17	0.10	-0.07	0.13
VII	11	2	18	0.77	0.18	0.19	1.11	0.18	0.10	0.00	0.24
VII	12	2	18	0.76	0.20	0.94	1.20	0.16	0.08	0.05	0.10
VIII	1	1	12	0.66	0.14	-0.19	0.30	0.16	0.14	0.03	0.02
VIII	2	2	24	0.68	0.13	-0.51	0.53	0.16	0.10	-0.08	0.35
VIII	3	2	24	0.70	0.14	-0.32	0.44	0.16	0.11	0.01	0.02
VIII	4	2	24	0.78	0.19	-0.16	0.62	0.19	0.16	-0.06	0.13
VIII	5	2	24	0.71	0.13	0.02	0.60	0.18	0.09	0.03	0.21
VIII	6	2	24	0.73	0.23	0.80	0.98	0.16	0.10	0.06	0.10
VIII	7	1	12	0.65	0.09	-0.22	0.38	0.16	0.08	0.03	0.02
VIII	8	2	24	0.72	0.21	-0.88	0.88	0.17	0.09	-0.09	0.28
VIII	9	2	24	0.73	0.23	-0.23	0.84	0.17	0.13	0.02	0.03
VIII	10	2	24	0.76	0.26	-0.11	0.92	0.18	0.11	-0.07	0.12
VIII	11	2	24	0.79	0.16	0.01	1.09	0.19	0.13	-0.05	0.23
VIII	12	2	24	0.81	0.30	0.79	1.50	0.17	0.11	0.00	0.13

Continued, next page

Table 3.7, cont'd:

Design	Subtest	Stage	No. Items	Mean a	SD a	Mean b	SD b	Mean c	SD c	Mean TIF Difference	MSE of TIF Difference
IX	1	1	24	0.65	0.13	-0.19	0.33	0.15	0.12	0.02	0.04
IX	2	2	12	0.71	0.16	-0.76	0.80	0.10	0.10	-0.18	0.36
IX	3	2	12	0.76	0.12	-0.32	0.52	0.19	0.13	0.06	0.04
IX	4	2	12	0.85	0.12	-0.13	0.51	0.23	0.11	-0.09	0.17
IX	5	2	12	0.75	0.12	0.18	0.63	0.17	0.12	0.06	0.26
IX	6	2	12	0.78	0.14	1.43	0.62	0.19	0.09	0.16	0.13
IX	7	1	24	0.67	0.12	-0.25	0.47	0.16	0.10	0.02	0.05
IX	8	2	12	0.73	0.18	-0.80	0.70	0.15	0.13	-0.14	0.31
IX	9	2	12	0.72	0.16	-0.37	0.60	0.14	0.09	0.02	0.03
IX	10	2	12	0.84	0.20	-0.23	0.54	0.21	0.12	-0.10	0.14
IX	11	2	12	0.71	0.14	0.32	0.62	0.15	0.12	0.05	0.24
IX	12	2	12	0.81	0.16	1.32	0.80	0.20	0.10	0.14	0.17

Continued, next page.

Table 3.7, cont'd:

Design	Subtest	Stage	No. Items	Mean a	SD a	Mean b	SD b	Mean c	SD c	Mean TIF Difference	MSE of TIF Difference
X	1	1	12	0.66	0.15	-0.20	0.29	0.16	0.13	0.04	0.02
X	2	2	12	0.67	0.10	-0.43	0.48	0.14	0.10	-0.10	0.20
X	3	2	12	0.70	0.10	-0.27	0.47	0.17	0.06	0.03	0.01
X	4	2	12	0.78	0.19	-0.29	0.56	0.22	0.14	-0.01	0.04
X	5	2	12	0.69	0.10	-0.01	0.53	0.17	0.06	0.04	0.10
X	6	2	12	0.66	0.13	0.73	0.86	0.14	0.10	0.08	0.07
X	7	3	12	0.73	0.21	-0.82	0.85	0.15	0.10	-0.14	0.19
X	8	3	12	0.71	0.24	-0.39	0.69	0.14	0.08	0.00	0.01
X	9	3	12	0.78	0.25	-0.28	0.84	0.18	0.16	-0.09	0.10
X	10	3	12	0.77	0.20	-0.02	1.13	0.17	0.13	-0.06	0.15
X	11	3	12	0.79	0.19	0.97	1.22	0.22	0.13	0.08	0.08
X	12	1	12	0.65	0.12	-0.24	0.38	0.16	0.12	0.02	0.02
X	13	2	12	0.67	0.17	-0.60	0.67	0.13	0.12	-0.11	0.18
X	14	2	12	0.70	0.18	-0.29	0.53	0.15	0.13	-0.01	0.01
X	15	2	12	0.79	0.19	-0.24	0.67	0.21	0.12	-0.04	0.05
X	16	2	12	0.76	0.16	-0.10	0.74	0.22	0.14	0.04	0.13
X	17	2	12	0.73	0.13	0.70	1.23	0.18	0.09	0.10	0.07
X	18	3	12	0.71	0.25	-0.93	0.91	0.17	0.11	-0.08	0.13
X	19	3	12	0.71	0.27	-0.35	1.13	0.15	0.06	0.01	0.01
X	20	3	12	0.72	0.28	-0.01	1.40	0.18	0.07	0.01	0.03
X	21	3	12	0.81	0.23	-0.26	1.48	0.16	0.09	-0.06	0.20
X	22	3	12	0.78	0.26	0.59	1.22	0.22	0.16	0.04	0.06

Continued, next page.

Table 3.7, cont'd

Design	Subtest	Stage	No. Items	Mean a	SD a	Mean b	SD b	Mean c	SD c	Mean TIF Difference	MSE of TIF Difference
XI	1	1	18	0.65	0.13	-0.22	0.34	0.15	0.12	0.03	0.03
XI	2	2	12	0.66	0.11	-0.50	0.79	0.10	0.09	-0.17	0.30
XI	3	2	12	0.70	0.12	-0.27	0.47	0.16	0.09	0.03	0.02
XI	4	2	12	0.81	0.19	-0.20	0.49	0.23	0.15	-0.04	0.08
XI	5	2	12	0.72	0.13	0.02	0.59	0.19	0.12	0.04	0.17
XI	6	2	12	0.78	0.26	1.13	0.84	0.19	0.11	0.09	0.06
XI	7	3	6	0.74	0.18	-1.01	0.97	0.17	0.13	-0.15	0.12
XI	8	3	6	0.72	0.16	-0.43	0.94	0.16	0.10	0.00	0.01
XI	9	3	6	0.91	0.23	-0.52	0.78	0.31	0.14	-0.01	0.03
XI	10	3	6	0.80	0.21	-0.21	1.09	0.20	0.08	0.00	0.13
XI	11	3	6	0.69	0.12	1.01	0.92	0.17	0.11	0.12	0.07
XI	12	1	18	0.65	0.08	-0.26	0.43	0.15	0.09	0.01	0.04
XI	13	2	12	0.69	0.18	-0.66	0.60	0.14	0.13	-0.13	0.26
XI	14	2	12	0.74	0.13	-0.44	0.70	0.17	0.11	-0.01	0.02
XI	15	2	12	0.78	0.27	-0.33	0.57	0.20	0.14	-0.03	0.06
XI	16	2	12	0.74	0.14	0.00	0.78	0.19	0.12	0.03	0.19
XI	17	2	12	0.78	0.20	1.01	1.00	0.19	0.08	0.09	0.08
XI	18	3	6	0.74	0.27	-1.30	0.86	0.20	0.09	-0.05	0.05
XI	19	3	6	0.74	0.23	-0.30	0.70	0.15	0.07	-0.01	0.01
XI	20	3	6	0.70	0.25	-0.03	0.74	0.17	0.08	-0.02	0.03
XI	21	3	6	0.84	0.18	-0.19	1.13	0.22	0.16	-0.03	0.12
XI	22	3	6	0.69	0.21	1.11	1.03	0.12	0.09	0.06	0.07

Continued, next page

Table 3.7, cont'd

Design	Subtest	Stage	No. Items	Mean a	SD a	Mean b	SD b	Mean c	SD c	Mean TIF Difference	MSE of TIF Difference
XII	1	1	6	0.64	0.08	-0.19	0.29	0.15	0.11	0.04	0.01
XII	2	2	12	0.64	0.06	-0.48	0.67	0.10	0.07	-0.12	0.11
XII	3	2	12	0.67	0.11	-0.25	0.33	0.14	0.10	0.01	0.01
XII	4	2	12	0.74	0.17	-0.21	0.49	0.21	0.13	0.00	0.03
XII	5	2	12	0.68	0.08	-0.07	0.51	0.16	0.07	0.02	0.08
XII	6	2	12	0.67	0.13	0.39	0.98	0.16	0.09	0.07	0.07
XII	7	3	18	0.71	0.20	-0.73	0.78	0.17	0.12	-0.06	0.15
XII	8	3	18	0.70	0.23	-0.23	0.73	0.15	0.10	0.02	0.01
XII	9	3	18	0.75	0.30	-0.18	0.59	0.18	0.11	-0.04	0.09
XII	10	3	18	0.76	0.16	-0.01	1.31	0.17	0.12	-0.03	0.15
XII	11	3	18	0.75	0.18	0.97	1.22	0.21	0.12	0.11	0.08
XII	12	1	6	0.63	0.10	-0.28	0.20	0.15	0.05	0.04	0.01
XII	13	2	12	0.70	0.19	-0.62	0.49	0.19	0.11	-0.02	0.08
XII	14	2	12	0.72	0.20	-0.22	0.67	0.16	0.18	0.00	0.01
XII	15	2	12	0.75	0.19	-0.16	0.64	0.16	0.14	-0.06	0.08
XII	16	2	12	0.79	0.17	-0.13	0.77	0.23	0.15	0.01	0.11
XII	17	2	12	0.72	0.19	0.35	1.03	0.18	0.11	0.07	0.10
XII	18	3	18	0.74	0.23	-0.67	1.22	0.17	0.10	-0.11	0.24
XII	19	3	18	0.78	0.28	-0.35	1.35	0.18	0.11	-0.04	0.03
XII	20	3	18	0.83	0.36	-0.20	1.16	0.18	0.10	-0.14	0.10
XII	21	3	18	0.77	0.21	0.10	1.20	0.15	0.11	-0.06	0.15
XII	22	3	18	0.83	0.22	0.48	1.32	0.21	0.10	0.00	0.15

3.4.1 Ability Estimation

There were three data analysis steps with respect to the precision of ability estimation of each test design. First, the accuracy of the ability estimates obtained from the 14 different test designs were assessed. Accuracy indicates the amount of error in the ability estimates. Since the true values of ability were known, the errors in ability estimation of each test design could be evaluated by comparing the estimates of ability to the true values. A measure of accuracy is the root mean squared error (RMSE) between the estimated and true ability values. For each test design, the $RMSE_a$ of ability estimates were calculated by computing the square root of the mean squared difference between the true and estimated ability for each examinee at each of the a (10) ability levels. $RMSE_a$ is given by:

$$RMSE_a = \sqrt{\frac{\sum_{j=1}^{n_a} (\hat{\theta}_j - \theta_j)^2}{n_a}}, \quad (3.4)$$

where $\hat{\theta}_j$ is the ability estimate of examinee j , θ_j is the true ability for examinee j , and n_a is the number of replications of estimates at ability level a .

In addition to calculating the RMSE to determine the amount of error in the ability estimates, the bias was calculated to determine whether the errors reflected a systematic tendency to overestimate or underestimate the ability. Bias is defined as the difference between the mean of the estimates and the true ability. For each test design, the $BIAS_a$ of ability estimates was calculated by computing the difference between the mean of the ability estimates and the true ability at each of 25 ability levels:

$$BIAS_a = (\theta_a - \bar{\hat{\theta}}_a), \quad (3.5)$$

where θ_a is the true ability at the a th ability level and $\bar{\hat{\theta}}_a$ is the average ability estimate at the a th ability level. Positive bias values indicate that ability was underestimated; negative values indicate that ability was overestimated.

Finally, the average standard error at each ability level for the 14 different test designs were calculated and compared graphically. To interpret the difference between average standard errors, the relative efficiencies of MST versus CAT and MST versus P&P were determined. The CAT and P&P designs were used as a baseline for comparing the efficiency of the MST designs. To determine the loss and gain of efficiency in MST designs as compared to CAT and P&P designs, respectively, the relative efficiency was computed by first calculating test information function at various ability levels for each MST design and comparing it to the test information function at the same ability level for the CAT and P&P designs. Test information is the sum of item information functions at θ and is given by:

$$I(\theta) = \sum_{i=1}^n I_i(\theta), \quad (3.6)$$

where $I_i(\theta)$ is item i 's information function at θ as described in equation 3.3 and n is the number of items in the test, in this case 36. (Hambleton, Swaminathan, & Rogers, 1991).

Relative efficiency (RE) is given by:

$$RE(\theta) = \frac{I_A(\theta)}{I_B(\theta)}, \quad (3.7)$$

where $I_A(\theta)$ and $I_B(\theta)$ are the test information functions for Tests A and B, respectively, defined over a common ability scale, θ (Hambleton et al, 1991). In this case, Test B corresponds to the CAT or P&P design and Test A corresponds to a MST design.

3.4.2 Item Exposure

For each of the 14 test designs, variable numbers of test items were used, and of the items used, they were administered to variable numbers of examinees. In this study, item exposure was analyzed by comparing the total number of items exposed and the number of people seeing each item in all 14 test designs. In addition, conditional exposure rates were compared for each test design.

CHAPTER 4

RESULTS

In this chapter, the results of the study are presented. According to the purpose of the study, the results are presented in two parts: precision of ability estimation and item exposure rates.

4.1 Ability Estimation

The precision of ability estimates obtained from all 14 test designs was analyzed in terms of accuracy, bias, and relative efficiency. First, the accuracy of ability estimates are presented and discussed.

4.1.1 Accuracy

The accuracy of the ability estimates obtained from the 14 different test designs was assessed using the root mean squared error (RMSE) between the estimated and true ability values. Figure 4.1 provides a line plot of the RMSEs at each ability level for the 14 test designs. Consistent with item pool difficulty and item information being centered around the middle of the ability distribution, the RMSEs were lowest in the middle region of the scale. The trend toward higher RMSEs at upper and lower ability levels is correspondingly due to there being fewer available items in the pool at the upper and lower regions of the scale (see Figure 4.2).

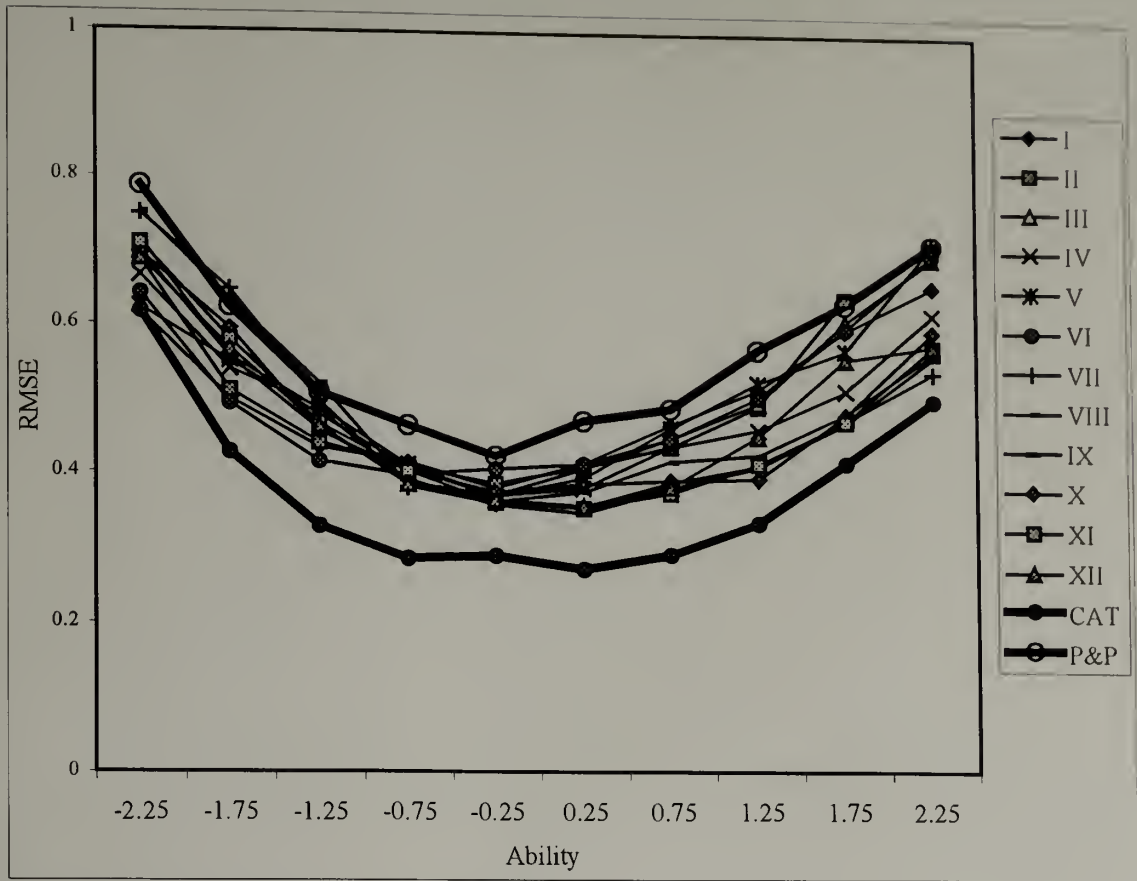


Figure 4.1. RMSEs of All 14 Test Designs

As shown in Figure 4.1, the CAT design led to less error in ability estimation (smaller RMSEs) at all ability levels than did any of the other test designs. This is not surprising since the CAT design adapted the difficulty of the test to the test taker's estimated ability after every item, rather than after administration of 6, 12, 18, or 24 items as in the MST designs or not at all as in the P&P design. As expected, the P&P design yielded the least accurate (highest RMSEs) ability estimates across all ability levels and the MST designs fell somewhere in between the P&P and CAT designs, in terms of accuracy of ability estimation.

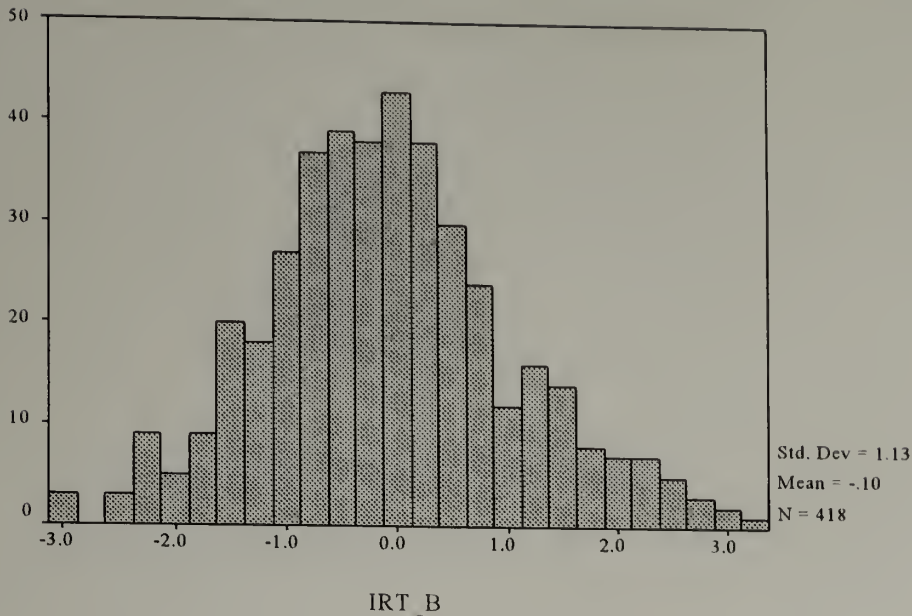


Figure 4.2. Frequency Distribution of b Parameters in the Pool

Figures 4.3 to 4.10 more clearly display the comparison between the 12 MST designs and the CAT and P&P designs. First, the accuracy of the two- and three-stage tests with three subtests in the second and third stages is described separately, followed by a comparison of the two- and three stage tests with three subtests. Secondly, the accuracy of the two-stage tests with five subtests in the second is described and a comparison of the accuracy of two-stage tests with three or five subtests in the second stage is made. Finally, the accuracy of three-stage tests with five subtests in the second and third stages is examined and the accuracy of the two- and three-stage tests with five subtests in the second and third stages are compared and the three-stage tests with three or five subtests in the second and third stages are compared.

Figure 4.3 displays the RMSE of the two-stage tests with three subtests at the second stage (MST I-III). In the middle and upper regions of the ability distribution (-.75 to 2.25), the three designs produced similar RMSE results, suggesting that varying the

number of items at each stage had little effect in the resulting accuracy of ability estimates in this region. However, in the lower end of the ability scale, varying the number of items per stage produced different results. Those designs employing unequal numbers of items per stage (MST II and III) led to more errors than did the MST design with equal numbers of items per stage (MST I).

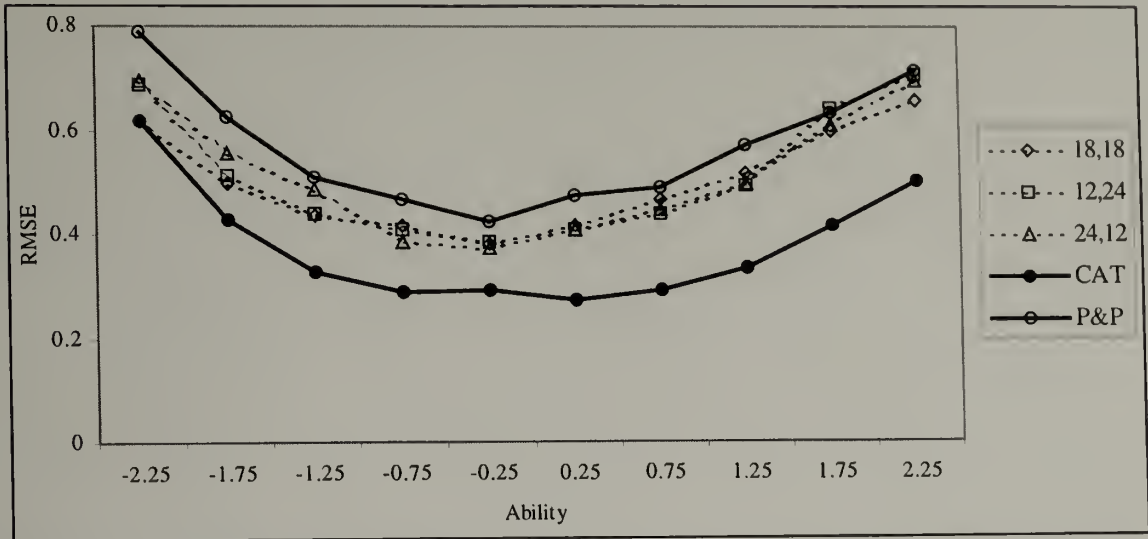


Figure 4.3. RMSEs of Two-Stage Tests with Three Subtests in the Second Stage (MST I-III)

The RMSEs from the three-stage tests with three subtests in the second stage are displayed in Figure 4.4 (MST IV-VI). As one might expect, increasing the number of stages from two to three increased the accuracy of ability estimation. At the higher ability levels (1.25 to 2.25), the RMSEs obtained from MST IV-VI were slightly smaller than those obtained from MST I-III (see Figure 4.5). Another finding showed that the three-stage test with three subtests in the second and third stages and equal numbers of items at each stage (MST IV) produced more accurate results than the other two MST designs. At the lower end of the ability scale, the MST design with the fewest number of items in the first stage (MST VI) led to more accurate ability estimates than did MST IV and V.

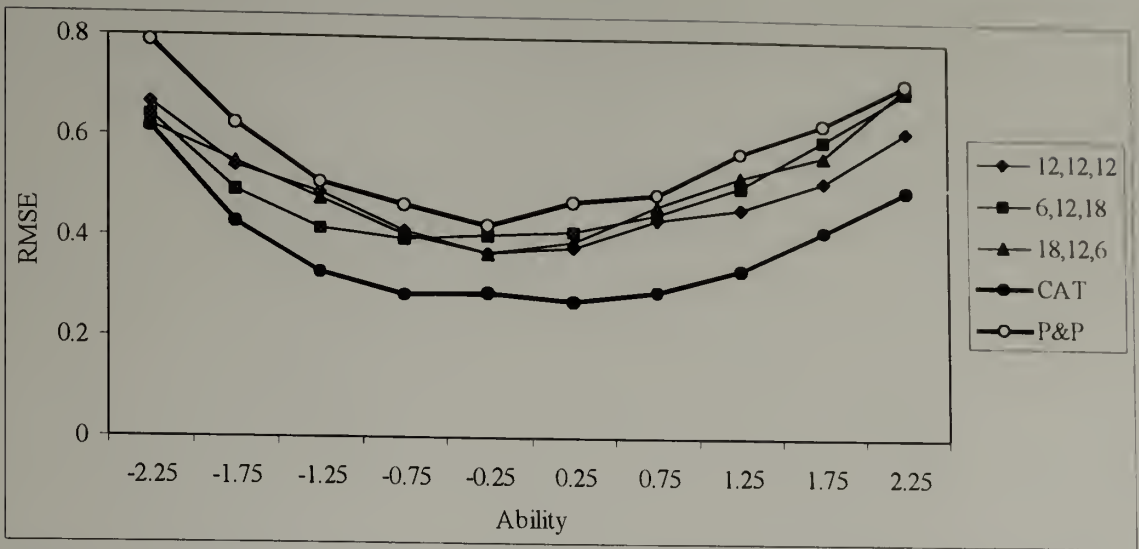


Figure 4.4. RMSEs of Three-Stage Tests with Three Subtests in the Second and Third Stages (MST IV-VI)

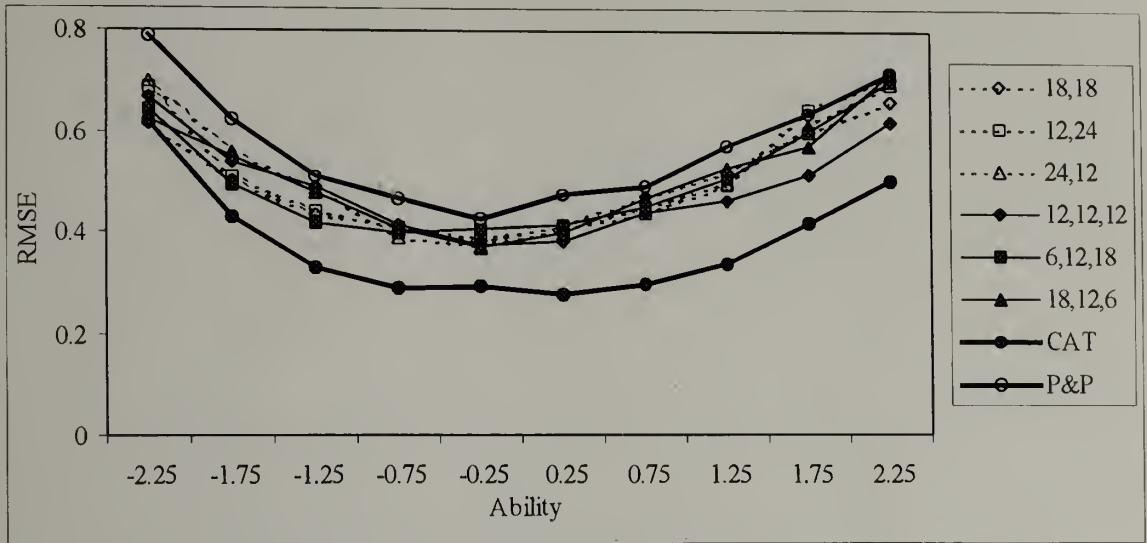


Figure 4.5. Comparison of Two- and Three-Stage Tests with Three Subtests in Second and Third Stages

Figure 4.6 displays the RMSEs for two-stage tests with five subtests in the second stage (MST VII-IX). In comparing the two-stage tests with three (MST I-III) and five (MST VII-IX) subtests in the second stage, it is apparent that there were fewer errors in ability estimates, especially at the higher ability levels, where there were five rather than

three subtests in the second stage (see Figure 4.7). In addition, as with MST I-III, there was little difference between varying the number of items in each subtest at each stage in the upper half of the ability scale for MST VII-IX. In the lower half of the ability scale, the designs with 18 and 24 items per subtest in the second stage (MST VIII and IX), yielded less accurate ability estimates than the design with only 12 items in each subtest in the second stage (MST VII).

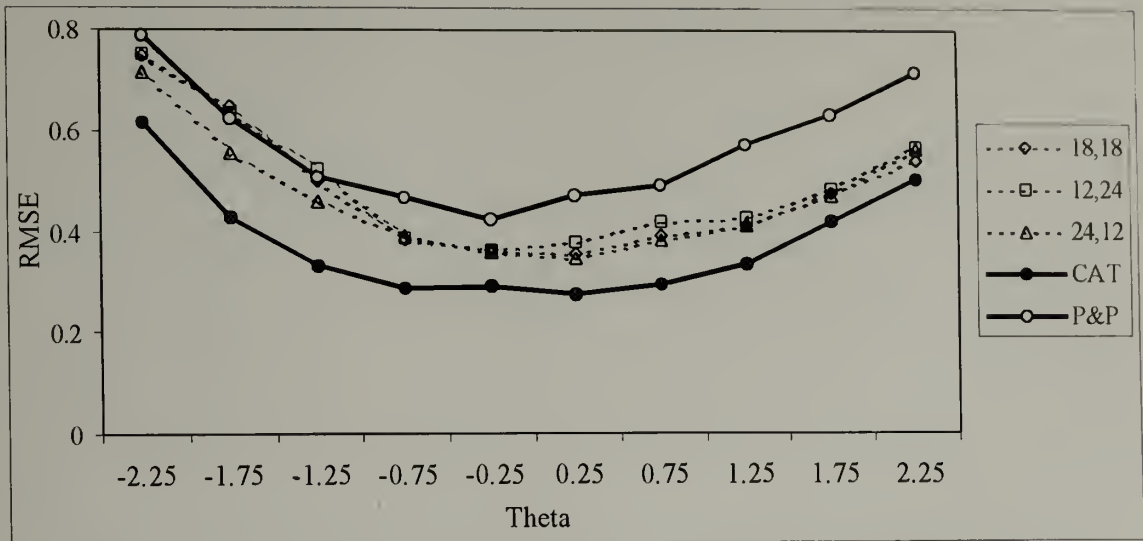


Figure 4.6. RMSEs of Two-Stage Tests with Five Subtests in the Second Stage (MST VII-IX)

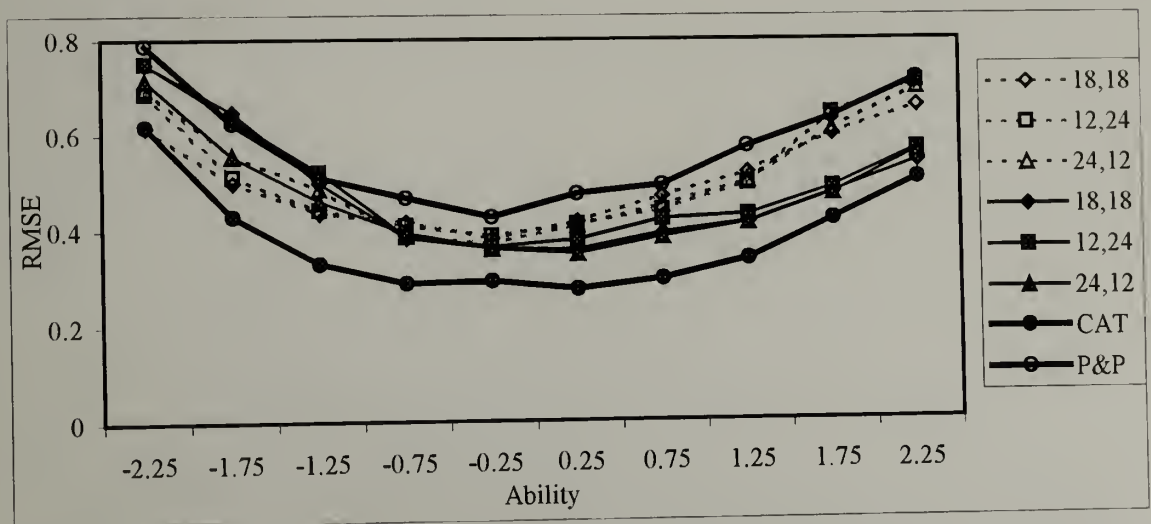


Figure 4.7. Comparison of Two-Stage Tests with Three or Five Subtests in the Second Stage

Finally, Figure 4.8 displays the RMSEs obtained from the three-stage tests with five subtests in the second and third stages (MST X-XII). With the exception of slightly smaller errors in the lower end of the ability scale in MST X-XII, there was little difference between MST VII-IX and MST X-XII, the two and three stage tests with five subtests in the second and third stages (see Figure 4.9). In comparing MST X-XII to MST IV-VI, the three-stage tests with five subtests versus three-stage tests with only three subtests, it is clear that the three-stage tests with five subtests in the second and third stages led to less error (see Figure 4.10). Finally, the number of items in each stage did not appear to affect the accuracy of ability estimates of MST X-XII.

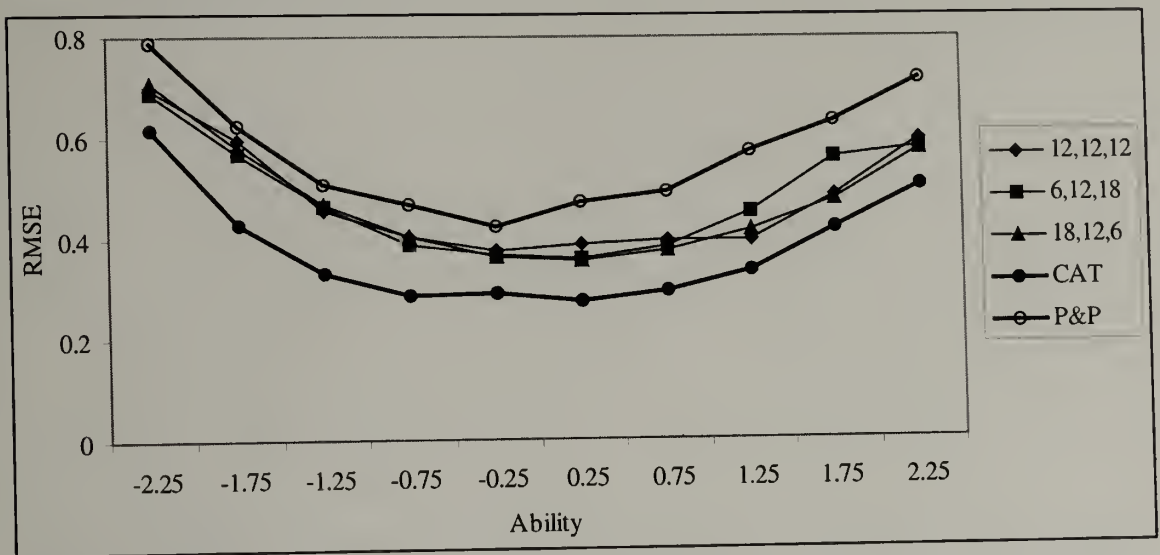


Figure 4.8. RMSEs of Three-Stage Tests with Five Subtests in the Second Stage (MST IX-XII)

Overall, MST VI, the three-stage test with three subtests in the second and third stages and 6, 12, and 18 items in the first, second, and third stages, respectively, produced the most accurate ability estimates at low ability levels (-2.25 to -1.25). At higher ability

levels (-.75 to 2.25), the multi-stage tests with five subtests in the second and third stages (MST VII-XII) led to the least amount of error in ability estimation.

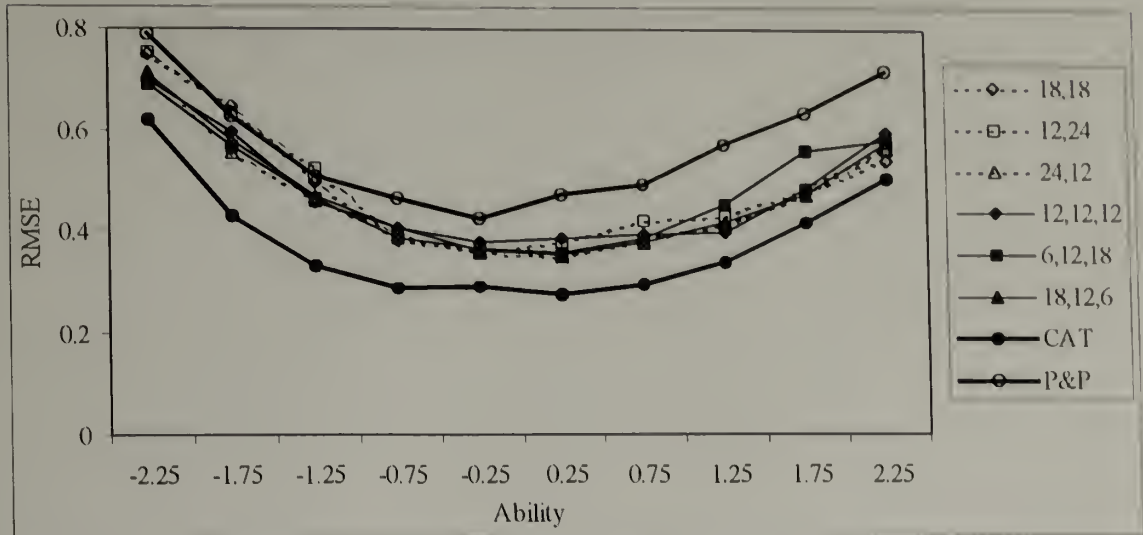


Figure 4.9. Comparison of Two- and Three-Stage Tests with Five Subtests in the Second and Third Stages

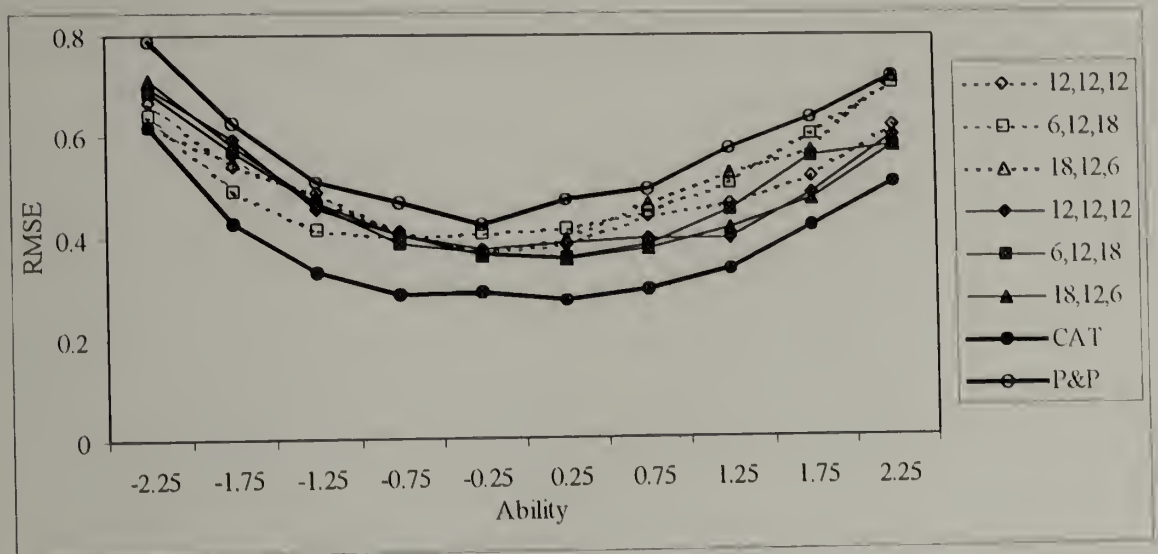


Figure 4.10. Comparison of Three-Stage Tests with Three or Five Subtests in the Second and Third Stages.

In summary, the three stage tests yielded lower errors in ability estimation than the two-stage tests. In addition, at most ability levels (-.75 to 2.25), increasing the number of

subtests from three to five increased the accuracy of ability estimation. Finally, at most ability levels (-.75 to 2.,25), varying the number of items per stage had little effect on the resulting accuracy of ability estimates, with the exception of the three-stage tests with three subtests in the second and third stages where the design with equal numbers of items per stage was superior.

4.1.2 Bias

As seen in Figure 4.11, there is a clear trend in the pattern of bias results across the ability scale for all 14 test designs. At the low end of the ability scale, the bias is positive, implying underestimation of the true ability values. The bias became negative as ability increased, implying an overestimation of examinees' true ability. Increased bias at the extremes of the ability scale was not an unexpected result. Rather, this reflects a well-known finding with maximum likelihood estimation (MLE) bias in the tails of the ability distribution (Lord, 1980).

Toward the upper tail of the ability distribution, the CAT design led to the least amount of bias in ability estimates. However, toward the lower tail of the distribution, many of the MST designs led to less bias in the ability estimates than either the CAT or P&P designs. The multi-stage tests that led to the least amount of bias in the ability estimates are the two- and three-stage tests with five subtests per stage (MST VII-XII).

The seriousness of the consequences of overestimating or underestimating an examinee's ability will vary by testing program. For instance, in medical licensing testing, one would assume that underestimating an examinee's ability level would be less severe

and risky in terms of consequences as compared to licensing an examinee as a medical doctor based upon an inflated ability estimate.

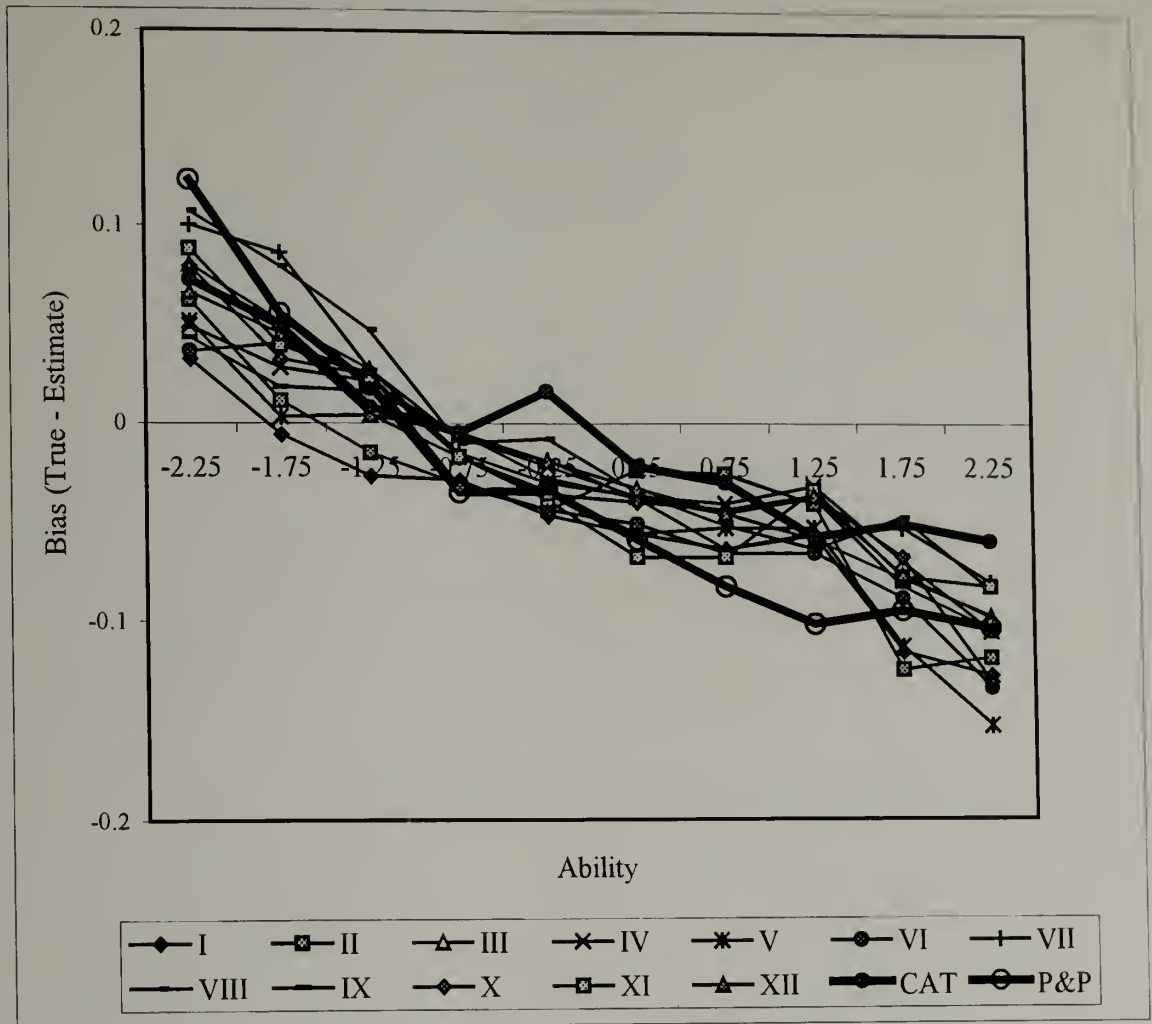


Figure 4.11. Bias of All 14 Test Designs

Overall, bias was small, -.12 to .13 and the number of subtests in the second and third stages and the number of items per stage had little effect on bias in the ability estimates.

4.1.3 Relative Efficiency

Tables 4.1 and 4.2 show the relative efficiency of each of the 12 MST designs to the CAT and P&P designs, respectively. As in describing the accuracy of ability estimates, the 12 MST designs are described in sets of three (I-III, IV-VI, VII-IX, and X-XII).

Figure 4.12 and Tables 4.1 and 4.2 report comparisons of the relative efficiency of the three two-stage tests with three subtests in the second stage (MST I-III) to both CAT and P&P. Overall, MST I-III tests were more efficient than the P&P test, but were not as efficient as the CAT. The efficiency in MST I-III was most apparent at the two extremes of the ability continuum with the two-stage test with 18 items at each stage (MST I) and was least apparent with MST III. At $\theta=2.25$, MST I-III tests were functioning as if they were 10-40% longer than the P&P test (see Table 4.1). That is, to yield the same precision of measurement as the multi-stage tests, the P&P test would need to be lengthened by adding 10-40% more comparable items to those items already in the test. Given the 36-item P&P test, this 10-40% translates into 4 to 15 items. This is not a trivial number of items when one considers the \$1500 cost of producing a single item. Relative to the CAT, at $\theta=2.25$, the multi-stage tests were 70% as efficient as the CAT, suggesting that the multi-stage tests would need to be lengthened by 140% (15 items) to produce estimates with the same precision as those produced by the CAT at the upper level.

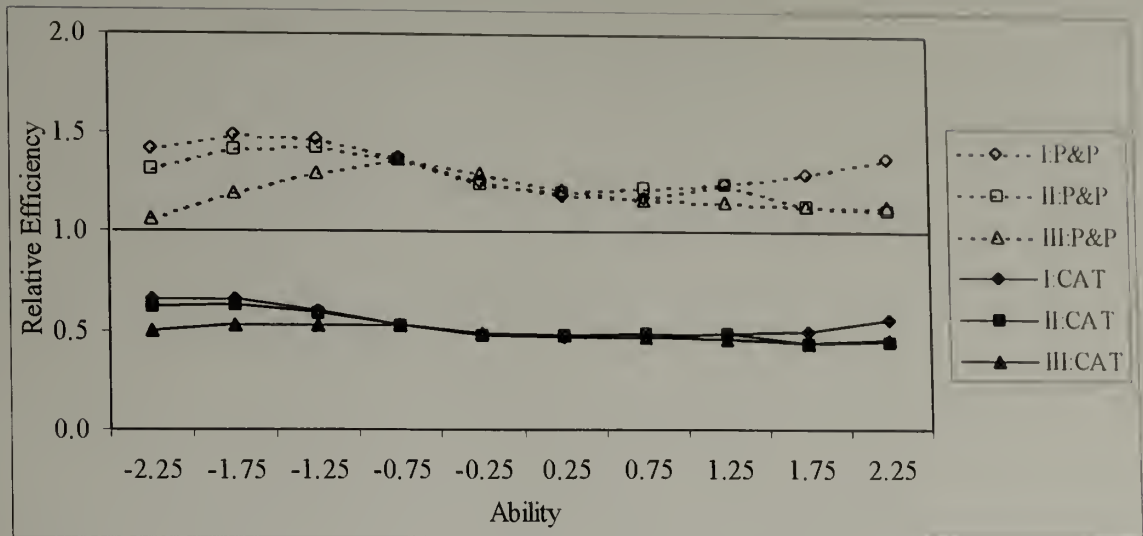


Figure 4.12. Efficiency of MST I-III Relative to CAT and P&P

Roughly the same pattern of relative efficiency that was observed for the two-stage tests with three subtests at the second stage (I-III) held for the three-stage tests with three subtests at the second and third stages (IV-VI; see Tables 4.1 and 4.2 and Figure 4.13). However, with MST IV-VI, the number of items per stage had less of an effect on the relative efficiency than it did with MST I-III. The greatest increase in efficiency for MST IV-VI was realized only in the lower end of the ability continuum.

Throughout most of the ability scale, the two-stage tests with five subtests in the second and third stages (MST VII-IX) showed greater efficiency over the P&P and CAT designs than did any of the two-stage tests (MST I-VI; see Tables 4.1 and 4.2 and Figure 4.13). As with MST IV-VI, varying the number of items in each stage had very little effect on efficiency throughout most of the ability range with MST VII-IX. However, in contrast to MST I-VI, a significant increase in efficiency was only attained at higher ability levels (1.25 to 2.25).

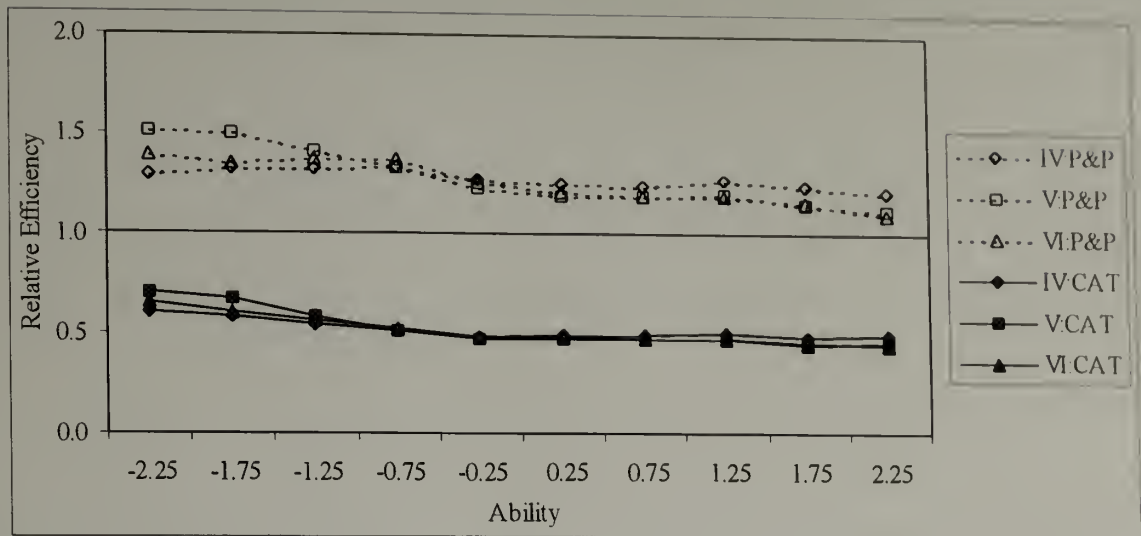


Figure 4.13. Efficiency of MST IV-VI Relative to CAT and P&P

Finally, the three-stage tests with five subtests in the second and third stages (X-XII) behaved similarly, in terms of relative efficiency, as did the two-stage tests with five subtests (MST VII-IX) in the second stage. Again, as with MST VII-IX, the number of items per stage had little effect on the relative efficiency of MST X-XII to P&P and CAT. Surprisingly, however, MST X-XII was slightly less efficient than MST VII-IX.

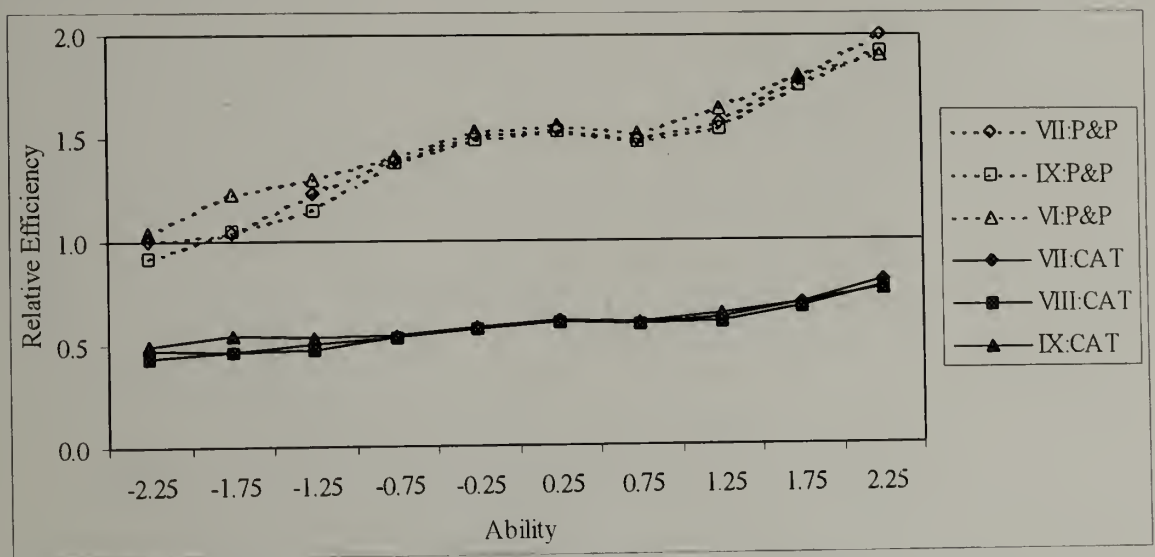


Figure 4.14. Efficiency of MST VII-IX Relative to CAT and P&P.

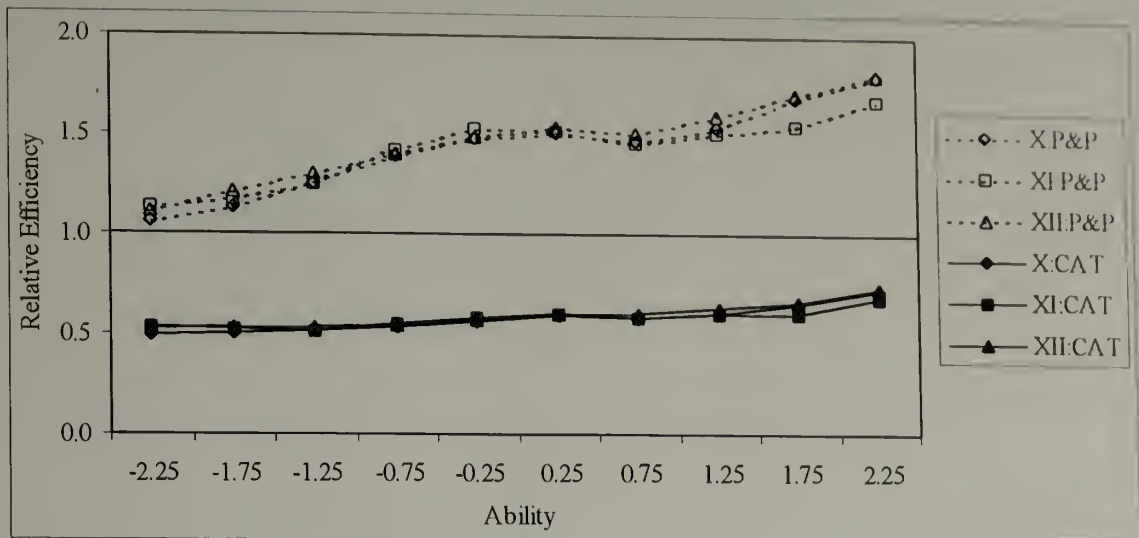


Figure 4.15. Efficiency of MST X-XII Relative to CAT and P&P.

Overall, MST I and VI were most efficient relative to both P&P (1.4-1.5 times) and CAT (.6-.7 times) at the lower ability levels (-2.25 to -1.25). At the higher ability levels, MST VII and VIII were most efficient relative to both P&P (1.5-2.0 times) and CAT (.6-.8 times).

In summary, increasing the number of stages from two to three had less of an effect on the relative efficiency of MST designs as compared to P&P and CAT than did increasing the number of subtests in the second and third stages from three to five. Furthermore, varying the number of items per stage evidenced little effect in relative efficiency of the MST designs as compared to the relative efficiency of the P&P and CAT designs.

Table 4.1

Efficiency of MST Relative to CAT

Ability	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII
-2.25	0.7	0.6	0.5	0.6	0.7	0.6	0.5	0.4	0.5	0.5	0.5	0.5
-1.75	0.7	0.6	0.5	0.6	0.7	0.6	0.5	0.5	0.5	0.5	0.5	0.5
-1.25	0.6	0.6	0.5	0.5	0.6	0.6	0.5	0.5	0.5	0.5	0.5	0.5
-0.75	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
-0.25	0.5	0.5	0.5	0.5	0.5	0.5	0.6	0.6	0.6	0.6	0.6	0.6
0.25	0.5	0.5	0.5	0.5	0.5	0.5	0.6	0.6	0.6	0.6	0.6	0.6
0.75	0.5	0.5	0.5	0.5	0.5	0.5	0.6	0.6	0.6	0.6	0.6	0.6
1.25	0.5	0.5	0.5	0.5	0.5	0.5	0.6	0.6	0.6	0.6	0.6	0.6
1.75	0.5	0.4	0.4	0.5	0.4	0.4	0.7	0.7	0.7	0.6	0.6	0.7
2.25	0.6	0.4	0.5	0.5	0.4	0.4	0.8	0.8	0.8	0.7	0.7	0.7

Table 4.2

Efficiency of MST Relative to P&P

Ability	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII
-2.25	1.4	1.3	1.1	1.3	1.5	1.4	1.0	0.9	1.0	1.1	1.1	1.1
-1.75	1.5	1.4	1.2	1.3	1.5	1.4	1.0	1.0	1.2	1.1	1.2	1.2
-1.25	1.5	1.4	1.3	1.3	1.4	1.4	1.2	1.1	1.3	1.3	1.2	1.3
-0.75	1.4	1.4	1.4	1.3	1.4	1.3	1.4	1.4	1.4	1.4	1.4	1.4
-0.25	1.3	1.2	1.3	1.3	1.3	1.2	1.5	1.5	1.5	1.5	1.5	1.5
0.25	1.2	1.2	1.2	1.2	1.2	1.2	1.5	1.5	1.6	1.5	1.5	1.5
0.75	1.2	1.2	1.2	1.2	1.2	1.2	1.5	1.5	1.5	1.5	1.5	1.5
1.25	1.2	1.2	1.1	1.3	1.2	1.2	1.6	1.5	1.6	1.5	1.5	1.6
1.75	1.3	1.1	1.1	1.2	1.1	1.2	1.8	1.8	1.8	1.7	1.5	1.7
2.25	1.4	1.1	1.1	1.2	1.1	1.1	2.0	1.9	1.9	1.8	1.7	1.8

4.1.4 Summary

All test designs showed very little bias and thus bias was not a dominant factor in comparing the MST test designs to the P&P and CAT designs. However, the test designs did differ in terms of their accuracy and relative efficiency. Essentially, the more branching that was done in the MST designs – the more stages and the more subtests per stage – the more closely did the results approximate the CAT design. In general, increasing the number of stages from two to three decreased the amount of errors in ability estimation. However, in some cases, it decreased the efficiency of the MST designs relative to P&P.

Increasing the number of subtests from three to five increased the accuracy of ability estimates as well as the efficiency of the MST designs relative to the P&P and CAT designs at most ability levels (-.75 to 2.25). Finally, at most ability levels (-.75 to 2.25), varying the number of items per stage had little effect on either the resulting accuracy of ability estimates or the relative efficiency of the MST designs to the P&P and CAT designs.

As noted in the Methods section, in reviewing the results, it is important to realize that the results are a function of the target information functions used to generate the P&P and MST designs. The explicit use of a target information function for assembling multi-stage tests places control over the amount and location of test precision in the hands of the test developer. The results are largely due to the target test information functions used. While efforts were taken to ensure that the comparisons between the P&P and MST and CAT and MST were fair, given these results, one could easily go back and increase the target information functions to reduce the RMSEs.

4.2 Item Exposure

Item exposure was examined in two steps. First, the number of items from the 418-item pool that were exposed in each test design was considered. Secondly, conditional item exposure rates were analyzed.

4.2.1 Number of Items Exposed

For each of the 14 test designs, variable numbers of test items were used, and of the items used, the items were administered to variable numbers of examinees. Table 4.3 contains all of this information. For example, with the second design in Table 4.3, the CAT design, 6% of the items were not administered to any of the 5000 examinees, 52% of the items were administered to between 1 and 500 examinees, 38% were administered to between 501 and 1000 examinees, and the remainder of the items (3%) were administered to between 1001 and 2000 examinees. The total percent of items NOT exposed and the total percent of items exposed are in boldface. Recall that five 36-item P&P forms and two panels of each MST design, constructed from a 418-item item pool, were used in this study.

Comparing the number of people seeing each item of the different test designs begs the question as to the importance of reducing the total number of items exposed OR reducing the number of examinees seeing each item. The two naturally compete against each other. While the philosophy of CAT tends to put more emphasis on the former, MST tends to favour the latter. Whereas the CAT design exposed 95% of the items with

Table 4.3

Percentage of Items Exposed to Different Numbers of Examinees in Each Test Design

Test Design	Total Percent NOT Exposed	Percentage of Items Exposed to the Given Number of Examinees						Total Percent Exposed
		1-500	501-1000	1001-2000	2001-3000	3001-4000	4001-5000	
P&P	57	0	26	17	0	0	0	43
CAT	6	52	38	3	0	0	0	94
I	66	9	9	9	9	0	0	34
II	60	11	0	23	6	0	0	40
III	71	3	9	6	11	0	0	29
IV	60	6	11	17	6	0	0	40
V	54	14	4	24	3	0	0	46
VI	66	9	6	11	9	0	0	34
VII	48	17	26	0	9	0	0	52
VIII	48	11	34	0	6	0	0	52
IX	60	11	17	0	11	0	0	40
X	45	11	37	0	6	0	0	55
XI	41	13	35	8	3	0	0	59
XII	50	16	24	1	9	0	0	50

a maximum 2000 people seeing any one item, the MST designs exposed only 29-60% of the items. The caveat is that with MST, some items were seen by up to 3000 examinees.

On the one hand, one may argue that once an item is seen by even one examinee, the item's security is compromised. Conversely, the argument can be made that if it is likely that many people are going to see the same item, the probability of someone seeing the same item is increased. Hence, the foreknowledge of items may play a bigger role in determining an examinee's ability estimate and thus the validity of test scores is called into question. In practice, it seems preferable to have many items exposed with each item being exposed to few people. Otherwise, there is no point to having the items in the bank.

As expected, CAT exposed the largest percentage of items (94%), with the majority (52%) of the items being exposed to between 1 and 500 people and no items being seen by more than 2000 examinees. For the MST designs, not surprisingly, the greater the number of stages, number of subtests per stage, and number of items in the higher stages, the greater was the number of items exposed. Of the MST designs, the three-stage test with five subtests in the second and third stages and 6, 12, and 18 items in the first, second, and third stages, respectively, (MST XI) exposed the largest number of items (59%). Finally, the P&P test exposed 43% of the 418 items with every item being seen by 500 to 2000 examinees.

Further investigation into the number of items exposed in the MST designs revealed that for some MST designs the number of items exposed was less than the number of items in the panel. Table 4.4 shows the number of items used to construct the five P&P forms and the two panels for each MST design. With the exception of four

MST designs, the percentage of items exposed equaled the number of items used to construct the five P&P forms and the two panels of each MST design. Interestingly, it was four of the six MST designs with five subtests in the second and third stages (MST VIII, X-XII) that exposed fewer items than were available. More specifically, Subtests 5 and 9 from MST VIII were never selected to be administered to any examinee; Subtests 3, 16, and 20 from MST X; Subtests 10, 14, 16, and 19 from MXT XI; and finally, Subtest 20 from MST XII was never administered (see Table 3.7). That some subtests were not administered from four of the six MST designs with five subtests in the second and third stages suggests that five subtests may have been too fine a distinction to route examinees.

Table 4.4

Number of Items Available and Number of Items Exposed

Test Design	Available	Exposed	Not Exposed
P&P	180	180	0
I	144	144	0
II	168	168	0
III	120	120	0
IV	168	168	0
V	192	192	0
VI	144	144	0
VII	216	216	0
VIII	264	216	48
IX	168	168	0
X	264	228	36
XI	312	252	60
XII	216	210	6

In summary, CAT exposed a large number of items (95%) with the majority of items (90%) being seen by less than 1000 examinees. On the other hand, MST exposed fewer items (29-60%) with many more people seeing each item. The MST design that exposed the most items was the three-stage test with five subtests in the second and third stages and 6, 12, and 18 items in the first, second, and third stages, respectively (MST XI).

4.2.2 Conditional Exposure Rates

Table 4.5 displays the conditional exposure rates and the average conditional exposure rate for each test design. Overall, the CAT and P&P designs yielded the comparable conditional exposure rates (.16 to .18). The MST designs yielded conditional exposure rates between .17 and .45. In general, increasing the number of stages from two to three and increasing the number of items in the first stage decreased the conditional exposure rates. Increasing the number of subtests from three to five had no systematic effect on the conditional exposure rates. Of the 12 MST designs, the three-stage test with five subtests in the second and third stages and 6, 12, and 18 items in the first, second, and third stages, respectively, (MST XI) yielded the smallest average conditional exposure rate (.27).

Recall that in constructing the multi-stage panels, a conditional exposure rate of .25 was used to specify the target information function for each subtest for two panels. However, aside from the decision to construct two multi-stage panels for each design, item exposure was not taken into account when simulating test-takers through a multi-stage test. In operational CAT programs that offer continuous testing with an item pool

Table 4.5
 Conditional Exposure Rates for Each Test Design

Test Design	Ability											Average
	-2.25	-1.75	-1.25	-0.75	-0.25	0.25	0.75	1.25	1.75	2.25		
P&P	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.16
CAT	0.17	0.17	0.17	0.18	0.17	0.16	0.17	0.17	0.17	0.17	0.17	0.17
I	0.42	0.42	0.42	0.42	0.28	0.23	0.28	0.32	0.42	0.42	0.42	0.36
II	0.42	0.42	0.42	0.42	0.26	0.19	0.35	0.42	0.42	0.42	0.42	0.37
III	0.43	0.43	0.43	0.43	0.33	0.31	0.32	0.32	0.39	0.43	0.43	0.38
IV	0.37	0.42	0.42	0.31	0.24	0.23	0.30	0.34	0.37	0.45	0.45	0.35
V	0.41	0.41	0.39	0.34	0.19	0.17	0.26	0.37	0.41	0.41	0.41	0.34
VI	0.42	0.42	0.42	0.42	0.27	0.20	0.28	0.39	0.43	0.45	0.45	0.37
VII	0.42	0.42	0.32	0.20	0.32	0.42	0.27	0.28	0.42	0.42	0.42	0.35
VIII	0.43	0.43	0.43	0.32	0.42	0.43	0.28	0.36	0.43	0.43	0.43	0.40
IX	0.42	0.31	0.23	0.21	0.26	0.35	0.21	0.21	0.38	0.42	0.42	0.30
X	0.34	0.31	0.28	0.24	0.23	0.35	0.24	0.22	0.38	0.42	0.42	0.30
XI	0.41	0.31	0.22	0.20	0.29	0.29	0.19	0.21	0.31	0.31	0.31	0.27
XII	0.42	0.43	0.42	0.26	0.32	0.34	0.26	0.31	0.42	0.42	0.42	0.36

being active for one week, conditional exposure rates greater than .35 are not acceptable. If one desired to do MST on a daily or continuous testing basis in the same fashion, one would want to make available more MST panels.

4.2.3 Summary

In summary, increasing the number of stages from two to three increased the number of items exposed and decreased conditional exposure rates. Increasing the number of subtests in the second and third stages increased the number of items exposed, but had little effect on the conditional exposure. Finally, increasing the number of items in the lower stages decreased the number of items exposed. However, conditional exposure rates even increased.

4.3 Summary

All test designs showed very little bias and thus bias was not a big factor in comparing the test designs. However, the test designs did differ in terms of their accuracy (RMSE) and relative efficiency. Essentially, the more branching that was done, the more likely were the results to approximate a computer-adaptive test. In general, increasing the number of stages from two to three decreased the amount of errors in ability estimation. However, in some cases, it decreased the efficiency of the MST designs relative to P&P.

Increasing the number of subtests from three to five increased the accuracy of ability estimates as well as the efficiency of the MST designs relative to the P&P and CAT designs, at most ability levels (-.75 to 2.25). Finally, at most ability levels (-.75 to 2.25), varying the number of items per stage had little effect on either the resulting accuracy of

ability estimates or the relative efficiency of the MST designs to the P&P and CAT designs.

CAT (which adapts at the item level rather than by stages) is impractical for many large-scale testing programs because of non-psychometric issues. For example, CAT is not practical for testing programs that have a large number of content specifications or for testing programs that must administer items that refer to a common stimulus (e.g., a reading passage). Additionally, item review in CAT is not common in practice because of the implication of mis-adapting a test to an examinee because he or she has changed their previous responses. From a test taker's perspective, item review would be welcomed enthusiastically. Furthermore, because CAT is built in real-time, there is no human intervention of quality assurance with the exception of what can be coded numerically.

Multi-stage testing provides a solution to these criticisms of CAT. Although the MST designs did not produce as accurate ability estimates as did CAT, nor were they as efficient as CAT, they do present themselves as a viable alternative to CAT as they provide item review within a stage, allow one to review the panels, and allow one to meet a large number of content constraints. In addition, the results from this study have shown that all MST designs have a positive impact on the quality of measurement over the result obtained with P&P tests of the same length

CHAPTER 5

CONCLUSION

5.1 Conclusion

There is considerable evidence demonstrating that computerized adaptive testing (CAT) and multi-stage testing (MST) are viable frameworks for testing. With many testing organizations looking to move toward CAT or MST, it was important to ascertain which framework functions best in different situations in terms of measurement accuracy and item exposure rates. What was needed was a systematic comparison of the different testing procedures under various realistic testing conditions. This dissertation addressed the paramount problems of the increase or decrease in accuracy of ability estimation and item exposure rates in using MST rather than CAT.

While there are many variables to consider when designing a multi-stage test, in this study, some variables were fixed to examine the effects of varying other variables on the accuracy of ability estimates and item exposure rates produced by MST. In this study, total test length was fixed, and the number of stages, the number of subtests per stage, and the number of items in each subtest were manipulated. The ability estimates and item exposure rates obtained from various designs of MST were compared with those obtained from CAT and a P&P test.

A simulation study was conducted using item parameters from a real item pool and ability parameters based on three-parameter logistic calibrations of real data. The primary question of interest was, given a fixed test length, how many stages and how many subtests per stage should there be in order to maximize measurement precision?

Furthermore, given a fixed test length, how many items should there be in each subtest? Should there be more in the routing test? Or should there be more in the higher stage tests? A secondary question of interest concerned conditional item exposure rates and the number of items exposed by CAT and the different MST designs.

Not surprisingly, the results of this study revealed that CAT produced more accurate ability estimates and lower conditional item exposure rates and was more efficient than any of the MST designs. However, given that it is not feasible for some testing organizations that have a grave need for quality control, the desire to offer item review to examinees, and many content constraints to implement CAT, results of this study also indicate that MST is an attractive alternative to CAT.

Within MST, there are many theoretical design issues to consider. This study considered three such issues – number of stages, number of subtests per stage, and number of items per stage. The interaction of the number of stages (2 or 3), the two levels of subtests (3 or 5), and the number of items per subtest (3 rationales) yielded 12 MST designs. Given these three issues, the number of MST designs examined in this study is by no means exhaustive. However, given the designs studied, what follows are recommendations for practice.

Based solely on accuracy and relative efficiency, results indicate that if there was interest primarily in the accuracy of estimating ability of able examinees, a two-stage test with five subtests in the second stage (MST VII, VIII, or IX) is preferred. The number of items per stage had little effect on the resulting accuracy of ability estimates or relative efficiency of the tests to P&P and CAT. Where there is interest in the lower half of the ability distribution, a three-stage test with three subtests in the second and third stages

and 18, 12, and 6 items in the first, second, and third stages, respectively, is preferred (MST VI).

Based solely on the results of item exposure, the two- and three-stage tests with five subtests in the second and third stages (MST VII, VIII, and X-XII) are the preferred MST designs, as they made the greatest use of the item pool. In terms of conditional exposure rates, the three-stage test with five subtests in the second and third stages and 6, 12, and 18 items in the first, second, and third stages, respectively, (MST XI) was best in terms of yielding the lowest conditional item exposure rates.

Combining the importance of accuracy of ability estimation, relative efficiency, and item exposure, the three-stage test with five subtests in the second and third stages and 6, 12, and 18 items in the first, second, and third stages, respectively, (MST XI) is preferred over the other MST designs.

CAT (which adapts at the item level rather than by stages) is impractical for many large-scale testing programs because of non-psychometric issues. For example, CAT is not practical for testing programs that have a large number of content specifications or for testing programs that must administer items that refer to a common stimulus (e.g., a reading passage). Additionally, item review in CAT is not common in practice because of the implication of mis-adapting a test to an examinee because he or she has changed his or her previous responses. From a test taker's perspective, item review would be welcomed enthusiastically. Furthermore, because CAT is built in real-time, there is no human intervention of quality assurance with the exception of what can be coded numerically.

Multi-stage testing provides a solution to these criticisms of CAT. Despite the fact that the MST designs did not produce equally accurate ability estimates as did the CAT design or that they were not as efficient as CAT, the MST designs present themselves as a viable alternative to CAT. MST provides item review within a stage, allows one to review the panels, and allows one to meet a large number of content constraints. In addition, the results from this study have shown that all MST designs have a positive impact on the quality of measurement over the result obtained with P&P tests of the same length.

5.2 Future Research

This study was limited to investigating the effects of two or three stages and three or five subtests per stage on the precision of measurement and item exposure rates. Given that there was some increase in precision of measurement and a decrease in conditional item exposure rates with three stages rather than two stages and five subtests rather than three, another study might investigate the number of stages and subtests it takes to approximate the precision of ability estimation in CAT and to lower item exposure rates to rates found in CAT.

Another logical extension of this study would be to examine the characteristics of the items that were highly exposed. It would be very beneficial to a testing program to be able to inform item writers of the popular item types (in terms of content and difficulty). Informing item writers would increase the richness of the item pool, while eliminating the cost associated with writing items that are rarely exposed and, thus, not needed.

Finally, the item selection algorithm, the way that content constraints and item exposure were managed, and how ability was estimated could be further investigated and manipulated. Throughout this study, several decisions were made in terms of how to weight information relative to content constraints and exposure rates, choosing a maximum exposure rate, choosing ability estimate bounds, and a standard error threshold. All of these decisions warrant further investigation.

REFERENCES

- Angoff, W. H., & Huddleston, E. M. (1968). The multi-level experiment. A study of a two-level test system for the College Board Scholastic Aptitude Test (Statistical Report No. 68-21). Princeton, NJ: Educational Testing Service.
- Clyman, S. G., Melnick, D. E., Clauser, B. E. (1995). Computer-based case simulations. In E. L. Mancall & P. G. Bashook (Eds.), Assessing clinical reasoning: The oral examination and alternative methods (pp. 139-149). Evanston, Illinois: American Board of Medical Specialties.
- Cronbach, L. J., & Gleser, G. C. (1965). Psychological tests and personnel decisions. Urbana, IL: University of Illinois Press.
- Federation of State Medical Boards of the U.S. & National Board of Medical Examiners (1998). Plans for administering the medical licensing examination on the computer: Special bulletin on computer-based testing (CBT) for the United States Medical Licensing Examination (USMLE). Philadelphia, PA: USMLE Secretariat.
- Graduate Management Admission Council (1997). GMAT Information Bulletin. Princeton, NJ: Educational Testing Service.
- Green, B. F. (1983). The promise of tailored tests. In H. W. Wainer and S. Messick (Eds.), Principals of modern psychological measurement (pp. 69-80). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. Newbury Park, CA: Sage.
- Hansen, D. N. (1969). An investigation of computer-based science testing. In R. C. Atkinson & H. A. Wilson (Eds.), Computer-assisted instruction: A book of readings (pp. 209-226). New York: Academic.
- Haynie, K. A., & Way, W. D. (1994, March). The effects of item pool depth on the accuracy of pass/fail decisions for the NCLEX using CAT. Paper presented at the meeting of the National Council on Measurement in Education.
- Kim, H., & Plake, B. S. (1993, April). Monte Carlo simulation comparison of two-stage testing and computerized adaptive testing. Paper presented at the meeting of the National Council on Measurement in Education, Atlanta, GA.
- Kingsbury, G. G., & Zara, A. R. (1989). Procedures for selecting items for in computerized adaptive tests. Applied Measurement in Education, 2, 359-375.

- Kingsbury, G. G., & Zara, A. R. (1991). A comparison of procedures for content-sensitive item selection in computerized adaptive tests. Applied Measurement in Education, 4, 241-261.
- Lord, F. M. (1952). A theory of test scores. Psychometric Monograph, No. 7.
- Lord, F. M., (1970). Some test theory for tailored testing. In W. H. Holtzman (Ed.), Computer-assisted instruction, testing, and guidance (pp. 139-183). New York: Harper and Row.
- Lord, F. M. (1971a). The self-scoring flexilevel test. Journal of Educational Measurement, 8, 147-151.
- Lord, F. M. (1971b). A theoretical study of two-stage testing. Psychometrika, 36, 227-242.
- Lord, F. M. (1977). A broad-range tailored test of verbal ability. Applied Psychological Measurement, 1, 95-100.
- Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Loyd, B. H. (1984, February). Efficiency and precision in two-stage adaptive testing. Paper presented at the meeting of the Eastern Educational Research Association, West Palm Beach, FL.
- Luecht, R. M. (1996). CASTISEL. Philadelphia, PA: National Board of Medical Examiners.
- Luecht, R. M., & Nungester, R. J. (1998). Some practical examples of computer adaptive sequential testing. Journal of Educational Measurement, 35, 229-249.
- Luecht, R. M., Nungester, R. J., & Hadadi, A. (1996, April). Heuristic-based CAT: Balancing item information, content and exposure. Paper presented at the meeting of the National Council of Measurement in Education, New York, NY.
- Lunz, M. E., & Bergstrom, B. A. (1994). An empirical study of computerized adaptive test administration conditions. Journal of Educational Measurement, 31, 251-263.
- Lunz, M. E., Bergstrom, B. A., & Wright, B. D. (1992). The effect of review on student ability and test efficiency for computerized adaptive tests. Applied Psychological Measurement, 16(1), 33-40.
- Mills, C. N., & Stocking, M. L. (1995, August). Practical issues in large-scale high-stakes computerized adaptive testing (Research Report 95-23). Princeton, NJ: Educational Testing Service.

- Olsen, J. B., Maynes, D. M., Slawson, D. A., & Ho, K. (1986, April). Comparison and equating of paper-administered, computer-administered and computerized adaptive tests of achievement. Paper presented at the meeting of the American Educational Research Association, San Francisco, CA.
- Patsula, L. N., & Steffen, M. (1997, March). Maintaining item and test security in a CAT environment. Paper presented at the meeting of the National Council of Measurement in Education, Chicago, IL.
- Robin, F. (1998). CATS: Computerized Adaptive Testing Simulation. Amherst, MA: University of Massachusetts, Laboratory of Psychometric and Evaluative Research.
- Rock, D. (1996). Two stage testing in ECLS. Unpublished manuscript.
- Rock, D. A., Pollack, J. M., & Quinn, P. (1995). Psychometric report for the NELS:88 base year through second follow-up (NCES 95-832). Washington, DC: U.S. Department of Education.
- Sands, W. A., Waters, B. K., & McBride, J. R. (1997). Computerized adaptive testing: From inquiry to operation. Washington, DC: American Psychological Association.
- Schaeffer, G. A., Steffen, M., Golub-Smith, M. L., Mills, C. N., & Durso, R. (1995). The introduction and comparability of the computer adaptive GRE General Test. Princeton, NJ: Educational Testing Service.
- Schnipke, D. L., & Reese, L. M. (1997, March). A comparison of testlet-based test designs for computerized adaptive testing. Paper presented at the meeting of the American Educational Research Association, Chicago, IL.
- Stocking, M. L. (1993). Controlling item exposure rates in a realistic adaptive testing paradigm. Princeton, NJ: Educational Testing Service.
- Stocking, M. L. (1996). Revising answers to items in computerized adaptive tests: a comparison of three models. Princeton, NJ: Educational Testing Service.
- Stocking, M. L., Lewis, C. (1998). Controlling item exposure conditional on ability in computerized adaptive testing. Journal of Educational and Behavioral Statistics, *23*(1), 57-75.
- Stocking, M. L., & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. Applied Psychological Measurement, *17*, 277-292.

- Stone, G. E., & Lunz, M. E. (1994). An investigation of procedures for computerized adaptive testing using partial credit scoring. Applied Measurement in Education, 7, 211-222.
- Van der Linden, W. J., & Boekkooi-Timminga, E. (1989). A maximin model for test design with practical constraints. Psychometrika, 54, 237-247.
- Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., Mislevy, F. J., Steinberg, L., & Thissen, D. (1990). Computerized adaptive testing: A primer. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Wald, A. (1947). Sequential testing. New York: Wiley.
- Weiss, D. J. (1973). The stratified adaptive computerized ability test (Research Report No. 73-3). Minneapolis, MN: University of Minnesota, Department of Psychology. Psychometric Methods Program.
- Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. Applied Psychological Measurement, 6, 473-492.
- Wise, S. L. (1996, April). A critical analysis of the arguments for and against item review in computerized adaptive testing. Paper presented at the meeting of the National Council on Measurement in Education, New York, NY.

