

1-1-2001

## The effect of test characteristics on aberrant response patterns in computer adaptive testing.

Saba M. Rizavi  
*University of Massachusetts Amherst*

Follow this and additional works at: [https://scholarworks.umass.edu/dissertations\\_1](https://scholarworks.umass.edu/dissertations_1)

---

### Recommended Citation

Rizavi, Saba M., "The effect of test characteristics on aberrant response patterns in computer adaptive testing." (2001). *Doctoral Dissertations 1896 - February 2014*. 5433.  
[https://scholarworks.umass.edu/dissertations\\_1/5433](https://scholarworks.umass.edu/dissertations_1/5433)

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](mailto:scholarworks@library.umass.edu).



UMASS/AMHERST

312066 0275 8398 1

THE EFFECT OF TEST CHARACTERISTICS ON ABERRANT RESPONSE  
PATTERNS IN COMPUTER ADAPTIVE TESTING

A Dissertation Presented

by

SABA M. RIZAVI

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

DOCTOR OF EDUCATION

September 2001

School of Education

© Copy right by Saba M. Rizavi 2001

All Rights Reserved

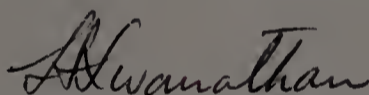
THE EFFECT OF TEST CHARACTERISTICS ON ABERRANT RESPONSE  
PATTERNS IN COMPUTER ADAPTIVE TESTING

A Dissertation Presented

by

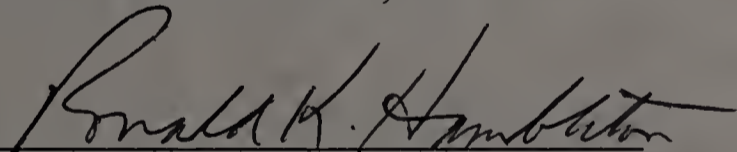
SABA M. RIZAVI

Approved as to style and content by:



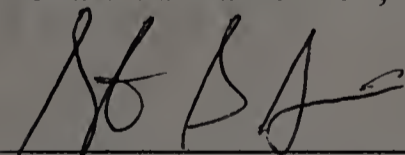
---

Hariharan Swaminathan, Chair



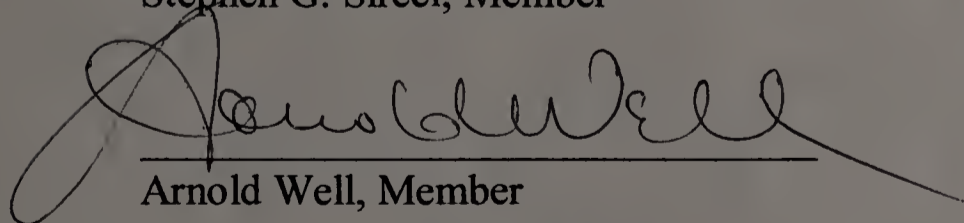
---

Ronald K. Hambleton, Member



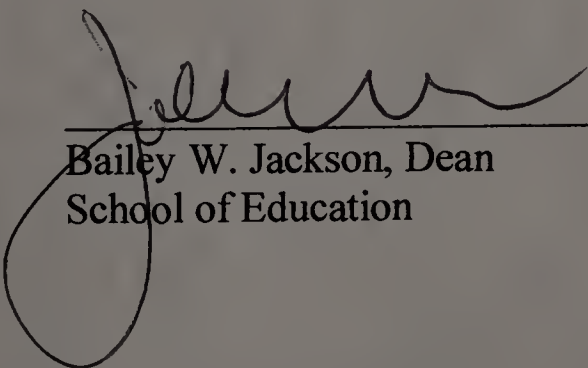
---

Stephen G. Sireci, Member



---

Arnold Well, Member



---

Bailey W. Jackson, Dean  
School of Education

DEDICATION

To my caring parents and loving husband

## ACKNOWLEDGEMENTS

I am pleased to express my sincere gratitude to many individuals for providing me with invaluable guidance and support throughout my program at University of Massachusetts at Amherst. First, I would like to thank my advisor Professor Hariharan Swaminathan for his guidance and care throughout my stay at University of Massachusetts. I am indebted to him for being understanding, supportive and kind at every point in my academic program. It was his generous direction and interest in my academic and personal life that I reached this final stage in the program. Second, I would like to express my deepest gratitude to Professor Ronald Hambleton for his continuous support and advice throughout the program and during this research. Professor Hambleton helped me develop into a successful professional in the field of research and psychometric methods. I thank him for going out of the way many times to provide me with useful feedback that enabled me to improve at every step of the program.

A special thanks goes to Professor Stephen G. Sireci who supported me during the course of study in every possible way. I thank him for giving me various opportunities to participate in professional activities in and out of the university environment. It was also his recognition of my strengths and weaknesses that helped me develop into a stronger person.

Sincere gratitude goes to my family, friends and colleagues for providing a friendly environment to aid my personal and professional development. I am truly appreciative of Dr. Charlena Seymour who provided me with courage and comfort and was always there for me in time of need. I am also very grateful to Dr. Frederic Robin whose advice was

invaluable while carrying out this research. It is also my pleasure to acknowledge the cooperation of Ms. Peg Lourainne who made things smooth at many instants in the course of study. It was her sweet personality that made me smile many times when I was feeling down.

I also want to take this opportunity to thank my husband for always being there for me in all my endeavors. It was his affection that soothed my mind whenever I was worried during the course of this research. I truly thank him for his devotion and for being so very accommodating to my work schedule and for his sweetness and patience.

I do not find enough words to thank my parents whose tender love and care will always be unforgettable. It is their efforts and support due to which I could achieve the best in my life. I cannot thank them enough for helping me deal with all hurdles in life to succeed and reach this level of education. Thank you Mom and Dad, I owe a successful life to you. /



## ABSTRACT

### EFFECT OF TEST CHARACTERISTICS ON ABERRANT RESPONSE PATTERNS IN COMPUTER ADAPTIVE TESTING

SEPTEMBER 2001

SABA M. RIZAVI, M.S. COMPUTER SCIENCE DEPARTMENT OF PESHAWAR  
UNIVERSITY, PAKISTAN

M.Ed., UNIVERSITY OF MASSACHUSETTS AMHERST

Ed.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Hariharan Swaminathan

The advantages that computer adaptive testing offers over linear tests have been well documented. The Computer Adaptive Test (CAT) design is more efficient than the Linear test design as fewer items are needed to estimate an examinee's proficiency to a desired level of precision. In the ideal situation, a CAT will result in examinees answering different number of items according to the stopping rule employed. Unfortunately, the realities of testing conditions have necessitated the imposition of time and minimum test length limits on CATs. Such constraints might place a burden on the CAT test taker resulting in aberrant response behaviors by some examinees. Occurrence of such response patterns results in inaccurate estimation of examinee proficiency levels. This study examined the effects of test lengths, time limits and the interaction of these factors with the examinee proficiency levels on the occurrence of aberrant response patterns.

The focus of the study was on the aberrant behaviors caused by rushed guessing due to restrictive time limits. Four different testing scenarios were examined; fixed length performance tests with and without content constraints, fixed length mastery tests and variable length mastery tests without content constraints. For each of these testing scenarios, the effect of two test lengths, five different timing conditions and the interaction between these factors with three ability levels on ability estimation were examined. For fixed and variable length mastery tests, decision accuracy was also looked at in addition to the estimation accuracy. Several indices were used to evaluate the estimation and decision accuracy for different testing conditions.

The results showed that changing time limits had a significant impact on the occurrence of aberrant response patterns conditional on ability. Increasing test length had negligible if not negative effect on ability estimation when rushed guessing occurred. In case of performance testing high ability examinees while in classification testing middle ability examinees suffered the most. The decision accuracy was considerably affected in case of variable length classification tests.

## TABLE OF CONTENTS

	Page
ACKNOWLEDGMENTS .....	v
ABSTRACT.....	vii
LIST OF TABLES .....	xii
LIST OF FIGURES .....	xiii
1. INTRODUCTION.....	1
1.1 Background .....	1
1.2 Statement of purpose .....	2
1.3 Scope .....	3
1.4 Outline of the study.....	4
2. AN OVERVIEW OF COMPUTER ADAPTIVE TESTING.....	5
2.1 Introduction of standardized tests .....	5
2.2 History of computer adaptive testing.....	5
2.3 Computer Based vs. Computer Adaptive tests.....	7
2.4 Item Response Theory .....	9
2.5 Features of Item Response Theory .....	10
2.5.1 Three-Parameter model.....	12
2.5.2 Two-Parameter model.....	14
2.5.3 One-Parameter model.....	14
2.6 Major components of a computer adaptive test .....	15
2.6.1 Item pool .....	15
2.6.2 Item response model .....	16
2.6.3 Starting point and initial estimate of ability. ....	18
2.6.4 Item selection strategy.....	19
2.6.5 Computation of the provisional estimate of ability .....	20
2.6.6 Termination criterion.....	22
2.6.7 Method for computing the final estimate of ability .....	23

3. CONSIDERATIONS IN THE DEVELOPMENT OF A CAT .....	25
3.1 Brief overview .....	25
3.1.1 Development of an item pool.....	25
3.1.2 Item exposure .....	26
3.1.3 Constrained Item Selection.....	29
3.1.4 Dimensionality .....	31
3.2 Examinee Interaction and test taking behaviors within a CAT environment.....	32
3.2.1 Examinee Interaction with a CAT.....	32
3.2.2 Examinee response times and test taking behavior .....	34
3.3 Aberrant response patterns.....	39
3.3.1 Definitions of an Aberrant Response in IRT framework.....	39
3.3.2 Appropriate Measurement or Person-fit research.....	39
3.3.3 Occurrence of aberrant response patterns in a CAT: .....	44
4. DESIGN AND METHODOLOGY.....	47
4.1 Introduction.....	47
4.2 Data Generation Model .....	48
4.3 Design.....	49
4.4 Item Pool characteristics using simulated parameters.....	51
4.5 Item Pool characteristics using AICPA parameters (without constraints) .....	52
4.6 Item Pool characteristics using AICPA parameters with content constraints.....	54
4.7 Mastery testing using AICPA parameters .....	56
4.8 Analyses.....	58
5. RESULTS .....	61
5.1 Results based on Simulations.....	61
5.2 Results for Proficiency Testing using AICPA parameters.....	65
5.3 Results for Proficiency Testing with Content Constraints using AICPA parameters .....	69
5.4 Results for Mastery Testing using AICPA parameters .....	70
6. CONCLUSION .....	79

APPENDICES

A. ABILITY AND ITEM PARAMETERS..... 83  
B. RESULTS USING SIMULATED ITEM PARAMETERS..... 86  
C. RESULTS USING AICPA ITEM PARAMETERS FOR AUDIT ..... 97  
D. RESULTS USING AICPA ITEM PARAMETERS FOR ARE ..... 124  
  
BIBLIOGRAPHY ..... 140

## LIST OF TABLES

Table	Page
2.1 Item Response Models.....	17
2.2 Ability and item parameter estimation procedures.....	21
4.1 Item parameter distribution.....	52
4.2 Ability parameter statistics.....	53
4.3 Item parameter statistics for Audit and Accounting & Reporting.....	54
4.4 Content specifications for Audit.....	55
4.5 Content specifications for Accounting & Reporting.....	56
4.6 Pool and test content composition.....	56
4.7 Derived cut-scores.....	58
5.1 Ability Levels.....	62
5.2 Error in estimates (RMSE) for Audit sub-test.....	66
5.3 Bias in estimates for Audit sub-test.....	66
5.4 Classification of masters/non-masters (Fixed Length Audit—75 items).....	70
5.5 Percentage of correctly classified at each ability level (Fixed Length Audit— 75 items).....	71
5.6 Classification of masters/non-masters (Fixed length Audit—30 items).....	72
5.7 Percentage of correctly classified at each ability level (Fixed length Audit— 30 items).....	73
5.8 Classification of masters/non-masters (Variable length Audit).....	74
5.9 Percentage of correctly classified at each ability level (Variable length Audit).....	76
5.10 Number of items taken by examinees at various ability levels (Variable length Audit).....	77

## LIST OF FIGURES

Figures	Page
1. Typical Item Characteristic Curve.....	13
2. Flowchart for a Computer Adaptive Test administration.....	24

# CHAPTER 1

## INTRODUCTION

### 1.1 Background

The advantages that computer adaptive testing offer over linear tests have been well documented. The computer adaptive test (CAT) design is more efficient than the linear test design in that with a CAT design fewer items are needed to estimate an examinee's proficiency level to a desired level of precision. This is accomplished by sequentially administering items that yield maximum precision at the examinee's current proficiency level. While this is highly desirable, a CAT with an item selection strategy that ignores such issues as content balance and exposure rates may compromise the validity of the test. Imposing content constraints and exposure controls on the CAT, while enhancing the validity of the test, imposes a considerable strain on the item pool and the administration of the CAT. A further issue that places a burden on the CAT test-taker is the imposition of limits on time and the minimum number of items that must be attempted. In the ideal situation, a CAT will result in different examinees answering different numbers of items according to the stopping rule employed. Imposing time limits is not employed in the ideal case, since the primary objective is to estimate an examinee's proficiency level with a desired level of precision. Unfortunately, the realities of testing conditions such as scheduling and improper test-taking strategies on the part of examinees, have necessitated the imposition of time and minimum test length limits. The constraints imposed on CATs that stem from validity-related issues as well as those based



on the realities of testing conditions may result in an examinee's proficiency level being estimated incorrectly.

Problems with estimating an examinee's proficiency level may occur for several reasons. An examinee may exhibit an aberrant response pattern such as responding correctly to a "difficult" item and incorrectly to an easy item. When such response patterns occur, especially with a three-parameter item response model, the likelihood function will not have a proper maximum, resulting in an inadmissible estimate for the examinee's proficiency level. Another problem in estimating examinee proficiency is that the item response model employed in CAT may not adequately model the examinee's performance. The existence of aberrant response patterns and the attendant inadmissible proficiency level estimates have been discovered in several testing programs.

## **1.2 Statement of Purpose**

The main purpose of the current study is to examine the reasons behind the occurrence of aberrant response patterns and their effect on the estimation of the proficiency level of an examinee. (It should be noted that this study is not aimed at examining procedures for detecting aberrant response patterns. Numerous procedures have been developed for detecting aberrant response patterns using appropriateness measurement indices and fit indices; these procedures will be used in identifying items that show aberrant response patterns in this study.) In order to explore the issue of aberrant response patterns, this study will examine the effect of time constraints on the occurrence of aberrant response patterns. The effect of interaction between the examinee proficiency level and various time constraints will also be studied. The study will also

examine the occurrence of such patterns in a variable length computer adaptive test designed for classification purposes.

An important factor in CAT is that of content constraints. The purpose of imposing content constraints is to enhance the validity of the test by ensuring that the content domain is represented. In fact, it can be argued that guessing on a relatively easy item by an examinee with high ability may be the result of content constraints; examinees with high ability value may not know a particular area of content and hence guess on an easy item from this content area. This will result in an aberrant response pattern. Thus, the existence of content constraints may provide an explanation of why aberrant response patterns occur.

While content constraints are not explicitly imposed in the study, the effects of content constraints can be studied to some degree from the proposed design (chapter 4). Since the effect of examinees guessing at various points on their response patterns can be interpreted from the content-constraint perspective, the net effect of imposing content constraints can be examined.

### **1.3 Scope**

Aberrant or non-model fitting responses occur when an examinee responds to test items in a manner that is not congruent with the underlying test model. This area of research has been known as appropriateness measurement in the past (Yi & Neiring, 1999; Drasgow & Levine, 1986) and as person-fit analysis more recently (Meijer & Neiring, 1995; Reise & Due, 1991). A variety of person-fit indices have been proposed to detect such aberrance and a great deal of research has been devoted to this issue (Bracey

& Rudner, 1992; Kogut, 1987; Kogut, 1986). While a variety of research has been conducted on detecting the extent of aberrance in a CAT, hardly any studies looked at the effect of test or examinee characteristics on the response pattern aberrance and how that aberrance in turn reflects on the ability estimates.

The present study looked at the issue of aberrance for fixed length achievement tests and both fixed and variable length mastery tests. Hence, the effect of test and examinee characteristics on the response aberrance was studied in the context of estimation accuracy as well as decision accuracy for adaptive tests designed and administered for a specific purpose. The research also considered the effect of aberrance on pool utilization and vice versa.

#### **1.4 Outline of the Study**

This research is organized into six chapters. In the first chapter, the background, the purpose, and the scope of the study have been considered. Chapter 2 discusses the history of computer adaptive testing and presents a description of the various components of a computer adaptive test. A pictorial representation of the computer adaptive testing process is included at the end of the chapter. Various considerations in the development of a CAT pool along with a discussion on response aberrance are presented in chapter 3. Chapter 4 describes in detail the design and methodology that was used to conduct the study followed by results of the research that constitute the next chapter. Finally, chapter 6 presents the main findings, limitations of the research and some future research directions.

## **CHAPTER 2**

### **AN OVERVIEW OF COMPUTER ADAPTIVE TESTING**

#### **2.1 Introduction of standardized tests**

The term “standardized test” generally means a group administered paper and pencil test in which everyone is tested under the same conditions. The idea is to measure each individual’s level of knowledge or achievement accurately for a desired purpose. During the earliest part of twentieth century, interest increased in the development of standardized tests from many different disciplines. Many new test formats that were aimed at large scale administration of such tests were being developed by researchers like Thorndike, Thurstone, Otis, and Terman (Carlson, 1994). As large scale testing became more popular a method was needed that could enable tests to be tailored to each person’s ability and knowledge thus reducing the time spent on the test while increasing the test efficiency at the same time. The intent of this chapter is to present a detailed overview of the computer adaptive testing in terms of history, theoretical background, and the various essential components and procedures involved in the implementation of such tests.

#### **2.2 History of computer adaptive testing**

The origin of computer adaptive testing can be traced back to the administration of Binet’s Intelligence Test in 1908, where the examiner chose the next question or task depending upon the examinee’s response on the last question or task administered (Hambleton, Swaminathan & Rogers, 1991; van der Linden, 1986; Weiss, 1983). The adaptive test brought to people’s attention the concept of tailoring the test to a test taker’s

ability by selecting the next item to be presented on the basis of performance on the preceding items (questions or tasks). After Binet's work, adaptive testing didn't gain much popularity due to the complexity of its implementation (Weiss, 1983).

The transition of adaptive testing from one-on-one to large group environment began in early 1950s when Hick (1951) initiated a testing program where the test "branched" into difficult or easy items depending upon the examinees' responses to previous items. The adaptive testing progressed through different stages of two-stage branching tests, pyramidal adaptive tests and stratified adaptive tests (for details, see Hambleton et al., 1991). In a two stage branching test, an examinee took a routing test and then took the next test (also called the "optimum test") based on the performance on the routing test. Unlike two stage tests, pyramidal and stratified adaptive tests involved multiple stages where examinees were presented with the same set of items but each examinee could take those items in a unique pattern. In the former type of test, the next item was the next available item with higher or lower difficulty level based upon the responses to previous items. The later, however, involved items that had been stratified into levels according to their difficulty level, hence, the next item was selected from these strata. (Weiss, 1982; Hambleton et al., 1991). The limitations of these tests were that not only these tests were mostly fixed length tests but also that they only considered item difficulty to create banks or pools of items while ignoring other characteristics like guessing and discrimination between various ability levels (Lord, 1980; Carlson, 1994).

In the late 1960s, Fred Lord initiated a comprehensive research program to pursue adaptive testing, focusing on the fact that fixed length tests could be replaced by variable length tests where the items were chosen in a way that they provided maximum

information about a person (Hambleton & Swaminathan, 1991). The advancement in computer technology made it possible for the test specialists to meet the implementation needs of adaptive testing and hence create adaptive tests. In US, the development of computer adaptive tests (CAT) was facilitated by research and conferences sponsored by Office of Naval Research and other agencies. Since the first conference on adaptive testing in 1975 (Clark, 1976), there has been an ongoing use of computer adaptive techniques to administer more efficient tests. Recently, CATs have become focus of testing agencies due to the range of benefits that they offer compared to the paper and pencil test. New forms of adaptive tests such as multistage adaptive tests (Patsula & Hambleton, 1999) where selectable entity is a mini-test instead of an item,  $\alpha$ -stratified multistage tests using item discrimination as a stratification factor (Chang, 1999), and ones using innovative models are under research.

### **2.3 Computer Based vs. Computer Adaptive Tests**

At this point it will be helpful if we clarify the two terms “computer based” and “computer adaptive”. Some applications of computerized testing use the computer only as a medium of presenting the test items. These tests are simply called computer-based tests (CBT) and they are no different from the paper and pencil tests except that they are administered on a computer. All of the examinees are administered the items in the same sequence as they would appear in the paper and pencil form of the test. These tests have advantages like rapid scoring, quick reporting and on-demand test delivery (ACE, 1995). However, since all the examinees are answering the same questions or items, there is no improvement in the test efficiency or precision. However, since CAT can be considered a

special kind of CBT, researchers often use the term “Linear on the Fly (LOFT)” test for a non-adaptive computer-based test.

Computer administration of adaptive tests, on the other hand, offers greater administrative standardization and much better techniques of item selection and scoring. Consider, for example a test designed to measure arithmetic ability at the fourth grade level. If the items were of appropriate difficulty for the average fourth grade level, students at this level would be measured with more precision: students above or below this level would be measured less precisely. If easier or harder items were added, it would not help if the range of difficulty were large. The easier items will be a waste of time for the able examinees and harder items may result in guessing thus affecting the measurement accuracy. Under a computer adaptive model, a computer program selects questions that target a candidate’s ability level; hence different examinees take different versions of the same test. Because of this matching of items to examinees’ abilities, a computer adaptive test is more efficient than a conventional paper and pencil or computer based test, usually requiring about half as many items to attain an equivalent precision in achieving the test goals (Weiss, 1982; Wainer et al., 1990). The relationship between “ability” and “getting an item right” has been highlighted well and in detail by the most widely used theory behind CAT, that is, item response theory (IRT). It is, therefore, essential to understand the importance of this theory in terms of its relationship with computerized adaptive testing. The next section of this chapter discusses in detail the Item Response Theory and its impact on adaptive testing.

## 2.4 Item Response Theory

Computer adaptive testing has matured to the point that it is now considered as an alternative to paper and pencil testing. Test publishers, licensure boards, private cooperation and school districts are beginning to implement CAT as an adjunct to or a replacement for their current paper and pencil tests.

Almost every application of CAT in the last 15 years has depended on the use of Item Response Theory (Kingsbury & Houser, 1993; Lord, 1980). In reality, the computer adaptive testing would not be feasible without item response theory (Hambleton, Swaminathan & Rogers, 1991). According to Weiss (1983), “when latent trait theories are applied to tests of ability or achievement, they have been known as item characteristic curve theory or most recently item response theory”. The latent trait theory is actually a set of models defining relationships between the observable variables and the underlying traits or constructs. The latent theory existed in one form or the other since early 1990s; the idea of this theory is actually implicit in classical test theory that had been used in testing since 1920s (Lord & Novick, 1968). The concept of true score in classical test theory and latent trait in latent trait theory were considered analogous to each other thus treating classical model as a simple case of a latent trait model.

It is difficult to trace the exact roots of latent theory, however, Mosier, Lawley and Guttman can be considered as the main contributors to the development of latent trait theory in mid forties. While Mosier (1941) was researching on test development theory and Lawley (1943) was focusing on the statistical aspects of the theory, Guttman (1944) developed the basics of latent trait theory to solve the scaling problems of attitude measurement (Weiss, 1983). Lord (1952) was the first person to actually apply the theory



When examinees rushed later in the test (after 75%), a slight drop was observed in the estimates for high ability examinees. The estimates for low and middle ability examinees generally remained stable when they guessed towards the very end. For some low and middle ability examinees, the estimates even improved when they guessed. The inaccuracy became evident when examinees started guessing after 75% of the test had been administered. The inaccuracy increased as the point in time at which guessing was introduced, moved earlier.

Increasing test length improved estimates by negligibly small amounts when guessing occurred later in the test. Increasing test length proved to have adverse effects on the estimation accuracy when the guessing started after almost 75-80% of the test had been administered. The negativity of increasing the test length was more significant for high ability examinees.

When the CAT was administered with content constraints, the error in estimation and the test information was not affected as long as enough items of varying difficulties existed in the pool. Same patterns of errors as well as test information were observed as in the case when the test didn't have content constraints. The average information was greatly affected when the pool lacked easy items in a content area and the examinees rushed to complete the test. In addition to the decrease in average information for that particular content area, a significant finding was the increase in average information for other content areas when guessing started early (except for very high ability examinees).

In case of mastery testing, the results were consequential. As guessing was introduced, people were incorrectly classified. The number of people who passed the test significantly decreased when guessing was introduced. The number of misclassifications

was larger for the variable length test compared to the fixed length test. The results indicated that examinees around the cut-point suffered most when guessing was introduced for both fixed and variable length tests. In case of fixed length CAT, the number of incorrectly classified examinees was larger for longer tests.

When the results from simulations using simulated item parameters were compared to those from simulations using AICPA parameters, it was found that the estimation for early guessers was better in the former case.

This study was limited in terms of several factors. The most significant limitation of the research was the imposition of time limits in the absence of a time-recording option.

The timing conditions can be simulated more accurately if the examinees' response times are actually recorded by a built-in timer or a clock. Another limitation was the simplicity of the test design in terms of test content, item formats and item types. The study focused on multiple-choice items only and represented major content strands on the test.

The adverse effects of the interaction of test and examinee characteristics with response aberrance can be reduced to some extent in several ways. Based on the results of the study, several suggestions can be made to address the issue. One suggestion is the use of a time/information index as introduced by Lou and Wang (2000). The time spent on an item could then be a part of the selection algorithm. Another possibility could be to build the aberrance flags into the weighted deviations model. In the other words, aberrant conditions could be controlled as part of the selection model.

In case of variable length mastery tests, increasing the minimum test length could prevent shorter tests to be administered to examinees. A minimum test length of 25 items

proved to be problematic and resulted in numerous false classification decisions when aberrance occurred.

The use of Bayesian estimation with stronger priors has proven to provide better estimates than Maximum Likelihood estimation as previously discussed in chapter 2. The same finding was reinforced by this research. The estimates were largely affected by aberrance when the rushed guessing was introduced early in the test. In real testing environment, rushed guessing is observed towards the later part of the test for most of examinees depicting aberrant response patterns. Hence Bayesian estimation can prove to be a better way to estimate ability for majority of the population. It is also expected that for early guessers, the MLE would lead to estimates much further from the truth compared to Bayesian estimates. Further research is however needed to shed light on this result.

It would be imperative to conclude by emphasizing on the well-stated fact that creating richer pools can always reduce the gravity of the problem. Frequently occurring aberrance could lead to unexpected utilization of the pool, hence it is highly desirable to have a large pool with informative items at all ability levels.

APPENDIX A

ABILITY AND ITEM PARAMETER DISTRIBUTIONS

estimates was large. The largest amount of bias was observed in high ability examinees while the smallest amount of bias was observed for low ability examinees. For low ability examinees, bias was higher for the shorter test when guessing was introduced later in the test. For medium and high ability examinees, difference in bias was negligible for the two test lengths when examinees guessed later. The differences, however, increased when examinees guessed earlier.

Figure C.4 and C.5 show the administration of a proficiency CAT to a typical low, medium and high ability examinee for Audit at two test lengths. The significant drop in the estimates for high ability examinees once they guessed early, explains the high values for RMSE. The accuracy was somewhat lost when examinees guessed towards the end; the loss was greater for high ability examinees. Another significant finding was that the estimates decreased significantly for middle ability examinees once they guessed after 75% of the test had been administered. The estimates decreased further when guessing began after half of the test was administered. The estimates, however, remained more stable compared to those for low and high ability examinees.

Figures C.6.a, and C.6.b represent the average test information at 12 ability levels for various guessing behaviors at two test lengths. As shown in figure C.6.b, the test provided maximum amount of information for middle to high ability examinees and minimum amount of information at the tails of the distribution. Similar pattern was observed when a shorter test was administered, however, as expected, the information was much lower than the 75-item test. The information stayed much more stable across the ability levels when compared with longer test in both guessing and non-guessing scenarios. When examinees guessed later in the test, the information was lost at most of

the ability levels except at higher ability levels. In case of ARE, a difference was that the information did not drop at the upper most tail of the distribution as was the case in Audit. When we look at figures C.7.a, and C.7.b for the average pool information at various ability levels, similar patterns were observed.

An interesting aspect of the study was to look at the average information that the pool provided before item selection algorithm began for each examinee. As mentioned earlier, the aberrance in examinee behaviors might have an adverse effect on the pool configuration. The selection of unusual number of easy or difficult items during the time when examinees rush into guessing could result in a less informative pool for various ability levels. For this purpose, Fisher's information was computed for each item in the pool. Information for each item was then summed to obtain the total amount of pool information that was available for item selection at the beginning of a CAT. In the previous section, reference was made to figures C.7.a, and C.7.b for pool information. Looking at the same figures, it was found that guessing also affected the amount of information that pool provided for item selection. An interesting finding was that early guessing resulted in a pool that provided maximum amount of information at the uppermost end of the ability distribution. This finding was specifically apparent when the examinees were administered a shorter test. An explanation for this might lie in the fact that exposure rates are subject to change significantly when examinees guess very early in the test. If each examinee guesses early, easier items must get utilized very quickly. At this point, it is useful to look at the pool utilization index. Figure C.8, depicts the plots for such index for two test lengths for several guessing scenarios. A slight increase was observed in the index when guessing increased. The value of the index was higher when

examinees were administered a shorter test indicating more Skewness in the exposure rates for items in the pool.

### **5.3 Results for Proficiency Testing with Content Constraints using AICPA Parameters**

The results for proficiency testing remained very similar when content constraints were introduced in the test. The results indicated that the content constraints did not seem to have much effect on the aberrance. The existence of sufficient number of easy items in the pool for each content area simplified the complexity of item selection that could arise when guessing was introduced. The plots for average test information for various content areas in Audit are depicted in figure C.12. Similar plots for a 30-item test are shown in figures C.13. The overall average test information is presented in figure C.14. The total test information for each examinee was re-scaled as the number of items in each content area was variable. The figures indicate that a large amount of information provided by the test was attributed to the first content strand.

In order to look at the effect of guessing when some content areas have fewer easier items than others, further simulations were conducted. The difficulty parameter for each item in the first content strand was increased by an arbitrarily chosen constant (1.2). The change in RMSE and Bias indices by changing the difficulty parameter for a single content strand was negligible, although that content area had the largest representation in the test. The plots for those indices are shown in Figure C.16.

The average test information for the various content strands is shown in figures C.17 and C.18. The average test information was significantly affected in the first content strand. A noticeable drop in the information was observed at all ability levels

except for examinees with abilities in the highest range. An interesting finding was that the information for other content areas increased for a wider ability range when examinees guessed earlier, even though configuration of the pool was not altered for those content areas. The information, in this case, decreased for the higher most ability levels.

#### 5.4 Results for Mastery Testing using AICPA Parameters

The classification decisions were first examined for the fixed length tests for Audit and ARE at the test lengths of 30 and 75 items. Table 5.4 presents the results for Audit for the test length of 75 items indicating the total number of people passed based on the true ability and then based on the estimated ability. The table also shows the overall percentage of people who were classified correctly versus those classified incorrectly as well as the percentage of misclassifications. Each of these results was then examined for

**Table 5.4: Classification of Masters/Non-Masters (Fixed Length Audit--75 items)**

Guessing points During CAT	People Passed		People Classified Correctly	People Class. Incorrectly	Percentage of Misclassifications
	True	Estimate			
No Guessing	400	371	1139	61	0.05
After 90%	400	277	1071	129	0.11
75%	400	149	949	251	0.21
50%	400	2	802	398	0.33
25%	400	0	800	400	0.33

the various guessing scenarios to look the effect of guessing on classification. Table 5.5 on the other hand shows the classification decisions broken down by ability level. The analyses indicated that out of 1200 examinees 400, examinees passed the Audit examination if the decisions were based upon their true abilities. The number of people



who passed the test was reduced to 371 when the decisions were based on the estimated abilities. When the numbers were broken down by ability, it was observed that

**Table 5.5: Percentage of Correctly Classified at each Ability Level (Fixed length Audit--75 items)**

Ability Levels	People Passed		Percentage of Correctly Classified				
	True	Estimate	No Guess	After 90%	After 75%	After 50%	After 25%
1	0	0	100	100	100	100	100
2	0	0	100	100	100	100	100
3	0	0	100	100	100	100	100
4	0	0	100	100	100	100	100
5	0	0	100	100	100	100	100
6	0	0	100	100	100	100	100
7	0	2	98	100	100	100	100
8	0	14	86	97	100	100	100
9	100	58	58	18	0	0	0
10	100	97	97	60	12	0	0
11	100	100	100	96	42	0	0
12	100	100	100	100	95	2	0

examinees only in the higher ability levels ( $\geq 0.54$ ) were able to pass the test. Based on the estimated ability, 16 out of 371 people who were originally classified in the middle ability levels were able to pass the test. The largest reduction in the number of people who passed the test based on the estimated ability was observed at the cut-point. In other words, the largest number of misclassifications was observed around the cut-point.

The results were then analyzed for various guessing behaviors. As shown in table 5.4, it was found that the number of examinees who passed the test, dropped by approximately 25% when the examinees started guessing after 90% of the items had been administered. The number dropped by 60% when the examinees started guessing after 75% of the items had been administered and by 99.5% when the examinees started guessing very early in the test. In terms of the classification decisions, the number of people that were misclassified when there was no guessing, doubled when examinees

started guessing after 90% of the items had been administered. The number of misclassifications increased by 4 times when guessing started after 75% of the test length and by approximately 6 to 7 times when examinees guess earlier. Out of the total population, the percentage of misclassifications increased from 5% to 11% for late guessing and 33% for earlier guessing.

As shown in table 5.5, the percentage of correctly classified increased from 98% to 100% for middle ability of 0.1 and from 86% to 97% for middle ability of 0.31 when examinees started guessing later in the test. However, the percentage significantly dropped (58% to 18%) for examinees at or slightly above cut-score. In other words, the accuracy of classification increased for examinees slightly below the cut-score once they started to guess. When the examinees started to guess earlier, the percentage of correct classifications dropped to 0% for examinees at or above cut-score.

Table 5.6 depicts the results from the similar analyses for Audit when the test length was reduced to 30 items. The overall number of people who passed the test with lesser number of items decreased. However as the examinees started to guess randomly at a certain point in the test, the number of examinees who passed the test increased when compared with the guessing behaviors in a longer CAT. For example, when examinees guessed after 75% of the 75-item test was administered, the number of people who passed

**Table 5.6: Classification of Masters/Non-Masters (Fixed Length Audit--30 items)**

Guessing points during CAT	People Passed		People Classified Correctly	People Class. Incorrectly	Percentage of Misclassifications
	True	Estimate			
No Guessing	400	366	1116	84	0.07
After 90%	400	270	1054	146	0.12
75%	400	192	986	214	0.18
50%	400	23	823	377	0.31
25%	400	1	801	399	0.33

the test was 149 compared to 192 people on a 30-item test. When guessing occurred after half way through the test, the number of examinees passing the test increased from 2 on a 75-item test to 23 on a 30-item test. The incorrect classifications were therefore more frequent in the case of a 75-item test when examinees started guessing relatively early in the test.

**Table 5.7: Percentage of Correctly Classified at each Ability Level (Fixed length Audit--30 items)**

Ability Levels	People Passed		Percentage of Correctly Classified				
	True	Estimate	No Guess	After 90%	After 75%	After 50%	After 25%
1	0	0	100	100	100	100	100
2	0	0	100	100	100	100	100
3	0	0	100	100	100	100	100
4	0	0	100	100	100	100	100
5	0	0	100	100	100	100	100
6	0	1	99	100	100	100	100
7	0	6	94	99	99	100	100
8	0	18	82	93	98	100	100
9	100	53	53	22	10	0	0
10	100	89	89	52	22	1	1
11	100	99	99	88	61	6	0
12	100	100	100	100	96	16	0

However, the situation was reversed when guessing occurred earlier. When broken down by ability levels, it was found that the decrease in the degree of misclassification or in other words increase in the percentage of correctly classified was observed for ability levels slightly below the cut-point. The degree of correct classification was decreased at ability levels at or slightly above the cut-point. As in the case of 75-item test, the classification decisions were accurate for very low and very high abilities.

When the examinees started to guess towards the end of the test, the differences between the proportions of correctly classified for late-guessing and no-guessing scenarios remained stable for the two test lengths. The only exception was observed for

very high ability examinees, where the proportion dropped by 11% in case of a 30-item test compared to 4% drop in the 75-item test. When guessing occurred at the beginning of the last quarter of the test (after 75%), the drop in the correct classification rates from no-guessing scenario was much larger for a longer test for high ability examinees. This was due to the fact that the number of correctly classified was higher for those examinees when they guessed earlier on the shorter test. For the examinees who were slightly above the cut-point, situation was similar to the 75-item test, that is, the percentage of correct classifications increased when examinees guessed later and became accurate when guessed earlier. The classification decisions for all examinees at or above the cut-score were inaccurate in both cases when guessing occurred very early in the test.

Table 5.8 indicates results for similar analyses when performed for a variable length test for Audit. As mentioned earlier, a stopping rule pertaining to the level of confidence in pass/fail decisions was employed in this case. The results were much

**Table 5.8: Classification of Masters/Non-Masters (Variable Length Audit)**

Guessing points during CAT	People Passed		People Classified Correctly	People Class. Incorrectly	Percentage of Misclassifications
	True	Estimate			
No Guessing	400	363	1133	67	0.06
After 90%	400	174	974	226	0.19
75%	400	85	885	315	0.26
50%	400	4	804	396	0.33
25%	400	1	801	399	0.33

similar to the fixed length when guessing occurred very early (after 25%) in the test or did not occur at all. Guessing at very early stage in the CAT resulted in highly inaccurate classification decisions for examinees at or above the cut-point in all cases. The results in this case, however, were much different from the fixed length test when examinees

guessed later in the test. In terms of the number of people who passed when there was no guessing involved, results were very similar for the fixed and variable length tests. When the examinees started to guess after 90% of items had been administered, the number of passing examinees dropped from 363 to 174, compared to a drop of 371 to 277 examinees on a 75-item test and a drop of 366 to 270 examinees on a 30-item test. The overall percentage of misclassifications was very similar for all testing modes for no-guessing situation (5% for 75 items, 7% for 30 items and 6% for variable length). The proportion of misclassified examinees increased from 11% on 75-item test and 12% on 30-item test to 19% on variable length CAT when examinees started guessing towards the end. Similarly, this proportion was increased from 18% on 75-item test and 21% on 30-item test to 26% on variable length test. When broken down by ability, it was observed that the accuracy of decisions suffered much more than the fixed length tests for examinees at or above the cut-score except those with the highest level of ability. The percentage of misclassification at cut-point decreased from 56% when the examinees did not guess, to 1% when guessing started after 90% of items had been administered. Similarly, these proportions decreased from 92% to 18% and 100% to 56% for the next higher levels of ability above the cut-score. The results for these two ability levels were rather drastic when compared to fixed length CATs. The drop in the proportion of correctly classified people in the above mentioned guessing scenario was approximately double for the examinees with ability level slightly higher the cut-point. The drop in accuracy was approximately five times the drop for a 30-item fixed length test and

**Table 5.9: Percentage of Correctly Classified at each Ability Level (Variable Length Audit)**

Ability Levels	People Passed		Percentage of Correctly Classified				
	True	Estimate	No Guess	After 90%	After 75%	After 50%	After 25%
1	0	0	100	100	100	100	100
2	0	0	100	100	100	100	100
3	0	0	100	100	100	100	100
4	0	0	100	100	100	100	100
5	0	0	100	100	100	100	100
6	0	0	100	100	100	100	100
7	0	3	97	100	100	100	100
8	0	12	88	100	100	100	100
9	100	56	56	1	0	0	0
10	100	92	92	18	3	0	0
11	100	100	100	56	21	3	1
12	100	100	100	99	61	1	0

approximately thirteen times for a 75-item test for examinees with abilities much higher than the cut-score.

It would be useful to look at the number of items that were attempted by the examinees before the test terminated. The following table shows the frequency of examinees that attempted a certain number of items. The numbers of items are grouped into five classes of 25, 26-30, 31-35, 36-40, 41-45. The results indicated that examinees around the cut-point attempted lesser items once they started guessing. It was also observed that the earlier they guessed, the lesser items they attempted. However, looking at table 5.10, the classification accuracy was adversely affected for slightly higher ability examinees above the cut-point and not for the ones below that threshold. In the first block of the table, a significant observation was that 67% of the examinees closest to the cut-point attempted 41-45 items. This proportion dropped to 4% when guessing was

**Table 5.10: Number of Items Taken by Examinees at Various Ability Levels  
(Variable length Audit Sub-test)**

	No Guess					Guessing Introduced after 90% of Items				
	25	26-30	31-35	36-40	41-45	25	26-30	31-35	36-40	41-45
1	100	0	0	0	0	100	0	0	0	0
2	100	0	0	0	0	100	0	0	0	0
3	100	0	0	0	0	99	1	0	0	0
4	100	0	0	0	0	100	0	0	0	0
5	98	2	0	0	0	96	3	1	0	0
6	86	4	5	2	3	93	5	2	0	0
7	64	7	5	6	18	84	11	5	0	0
8	38	11	7	4	40	54	36	9	0	1
9	12	13	5	3	67	26	43	21	6	4
10	41	12	10	4	33	19	38	33	10	0
11	85	10	3	0	2	56	10	27	5	2
12	100	0	0	0	0	99	0	1	0	0
<b>Total</b>	<b>924</b>	<b>59</b>	<b>35</b>	<b>19</b>	<b>163</b>	<b>926</b>	<b>147</b>	<b>99</b>	<b>21</b>	<b>7</b>
<b>Guessing Introduced after 75% of Items</b>										
	25	26-30	31-35	36-40	41-45	25	26-30	31-35	36-40	41-45
1	100	0	0	0	0	100	0	0	0	0
2	100	0	0	0	0	100	0	0	0	0
3	100	0	0	0	0	100	0	0	0	0
4	100	0	0	0	0	100	0	0	0	0
5	100	0	0	0	0	99	1	0	0	0
6	100	0	0	0	0	99	1	0	0	0
7	93	7	0	0	0	98	2	0	0	0
8	85	12	3	0	0	96	4	0	0	0
9	68	17	15	0	0	94	6	0	0	0
10	43	36	16	4	1	91	6	2	1	0
11	32	37	15	13	3	76	20	4	0	0
12	58	19	17	5	1	54	31	12	3	0
<b>Total</b>	<b>979</b>	<b>128</b>	<b>66</b>	<b>22</b>	<b>5</b>	<b>1107</b>	<b>71</b>	<b>18</b>	<b>4</b>	<b>0</b>
<b>Guessing Introduced after 25% of Items</b>										
	25	26-30	31-35	36-40	41-45					
1	100	0	0	0	0					
2	100	0	0	0	0					
3	100	0	0	0	0					
4	99	1	0	0	0					
5	100	0	0	0	0					
6	100	0	0	0	0					
7	100	0	0	0	0					
8	99	1	0	0	0					
9	100	0	0	0	0					
10	99	0	1	0	0					
11	98	2	0	0	0					
12	97	2	1	0	0					
<b>Total</b>	<b>1192</b>	<b>6</b>	<b>2</b>	<b>0</b>	<b>0</b>					

introduced towards the end of the test. A large number of those examinees attempted 26-30 items.

The next phase of the analyses were performed to look at the estimation accuracy of the fixed and variable length mastery tests. As expected, the estimation accuracy remained very similar to the proficiency testing for both fixed and variable length test. The results of mastery testing are presented in figures C.19 to C.24. An interesting fact was observed when we looked at the average information for a variable length test. The information monotonically increased till it peaked for examinees around the cut-score. After that point it became increasingly less till the uppermost ability level. This indicates that the examinees with abilities around the cut-score were presented most informative items. The information, in general, was decreased when compared with the fixed length mastery tests, being closer to the information provided by 30-item test. Interestingly, the peak disappeared when the examinees started guessing. The information curve remained relatively flat over the ability levels when examinees guessed. As mentioned above, a large number of examinees took longer test when guessing was not introduced, while that number significantly decreased when guessing was introduced. The average pool information, however, followed a pattern very similar to a 30-item fixed length test.



## CHAPTER 6

### CONCLUSION

The study shed light on some of the most important issues in computerized adaptive testing. The purpose of any assessment instrument is not if the information obtained on that instrument leads to an incorrect decision. In adaptive testing, an incorrect decision at any point in the test can lead to serious discrepancies towards the end. Since each item or question that gets administered to an examinee has impact on the properties of the remainder of the test, any disruption in the test administration process is consequential.

The act of an examinee rushing into random guessing at any point in the test could result in misleading estimates of that examinee's proficiency. As serious as it is in proficiency or achievement tests, the problem of inaccurate estimation could be worse in Mastery testing. The declaration of examinees as masters or non-masters on the basis of incorrect measures of their ability is no-doubt harmful.

The results of the study clearly indicate that the error in estimation increases significantly once the examinee rushes to finish the test. One could be misled into assuming that the low level examinees would be affected most by disruption in the item selection algorithm. The results of the study showed that the high ability examinees suffered most once they ran out of time. In all cases, the error in estimates was lowest for low ability examinees, higher for middle ability examinees and highest for high ability examinees.

to ability and achievement testing environment (using dichotomous items) thus giving shape to item response theory. The item response theory gained researchers' attention in no time resulting in an extensive research by many scientists like Lord and Novick (1968) and also in Europe by people like Rasch (1960) who tried to refine the models specified by item response theory. The practical implementation of item response theory was made feasible through the advent of computers in the late 1960s and that is when item response theory and computer adaptive testing merged (Weiss, 1983). The mechanical branching rules to select items were in most cases replaced by the item response theory procedures and since then item response theory (IRT) dominates the computer adaptive testing.

## 2.5 Features of Item Response Theory

It is desirable at this point to understand the logic behind the use of item response theory as the underlying theory behind computer adaptive testing. According to Hambleton and Swaminathan (1985),

These models such as the classical test model, are based upon weak assumptions, that is, the assumptions can be met easily by most test data sets, and therefore, the models can and have been applied to a wide variety of test development and test score analysis problems.....The purpose of any test theory is to describe how inferences from examinee item responses and /or test scores can be made about unobservable examinee characteristics or traits that are measured by a test. Presently, perhaps the most popular set of constructs, models, and assumptions for inferring traits is organized around latent trait theory.... or item response theory as Lord (1980) preferred to call the theory.

Hence the basic idea behind item response theory is that the test score or test performance of an individual can be described by one dominant factor among other factors that can effect performance, most commonly known as *ability*. If we plot the traits for various examinees against their performance on various tasks, questions or items, we will get a monotonically increasing function in most cases. In case of item

response theory where we assume ability as the only prominent factor, this function is called item characteristic function. Each item also has an item information function measured on the same scale as the ability scale which gives the information provided by an item at a certain point on the ability scale (Hambleton, Zaal & Peters, 1991). This feature makes item response theory most feasible for use in computer adaptive testing as the item can be selected for administration depending upon the amount of information it gives at a certain ability level.

The three main features of item response theory that make it useful over classical test theory are as follows, assuming ability being the only trait that the items are measuring:

- The examinee performance, or in more technical language the estimates of ability, are independent of the sample of items that are administered to them
- The item characteristics or in other words estimates of item parameters are independent of the sample of examinees taking those items
- It is possible to find the standard error of measurement for each ability estimate. This is a more useful way of determining precision of measurement since it allows for the fact that precision may be higher for certain values of ability (Hambleton, 1983).

Another distinctive feature of IRT is the different kinds of item response models it supports. An item response model depends upon the kind of data to be analyzed. The response data produced from a test can be dichotomous, polytomous or continuous. The dichotomous response pattern consists of only two categories (generally 'correct' or 'incorrect') while the polytomous response pattern consists of three or more categories. The continuous response pattern, on the other hand, consists of a very large number of

categories. An item response model also depends upon the shape of the distribution of the measurement errors. For example, in case of multiple choice dichotomous data, if the errors are normally distributed, normal-ogive models are used; whereas for logistically distributed errors, logistic models are used. The logistic and normal-ogive distributions are very similar; however, the logistic models are mathematically easier to work with. Traditionally adaptive tests have taken advantage of logistic IRT models (Kingsbury & Houser, 1993), however, recently researchers like Dodd (1990) have discussed the use of expanded item response models. The three logistic models are explained as follows:

### 2.5.1 Three-Parameter Model

The most commonly used item response model used for adaptive testing is the three parameter model which accounts for the level of difficulty for items or tasks, the extent to which the items discriminate among people and the amount of guessing to reach the correct response if a person is unable to answer correctly (Birnbaum, 1968). The exponential form of the 3-pl model is given as,

$$P_i(\theta) = c_i + (1 - c_i) \frac{\exp(Da_i(\theta - b_i))}{1 + \exp(Da_i(\theta - b_i))}$$

where,

$P_i(\theta)$  = probability of a correct response to an item  $i$  at an ability level  $\theta$  ,

commonly known as Probability Function

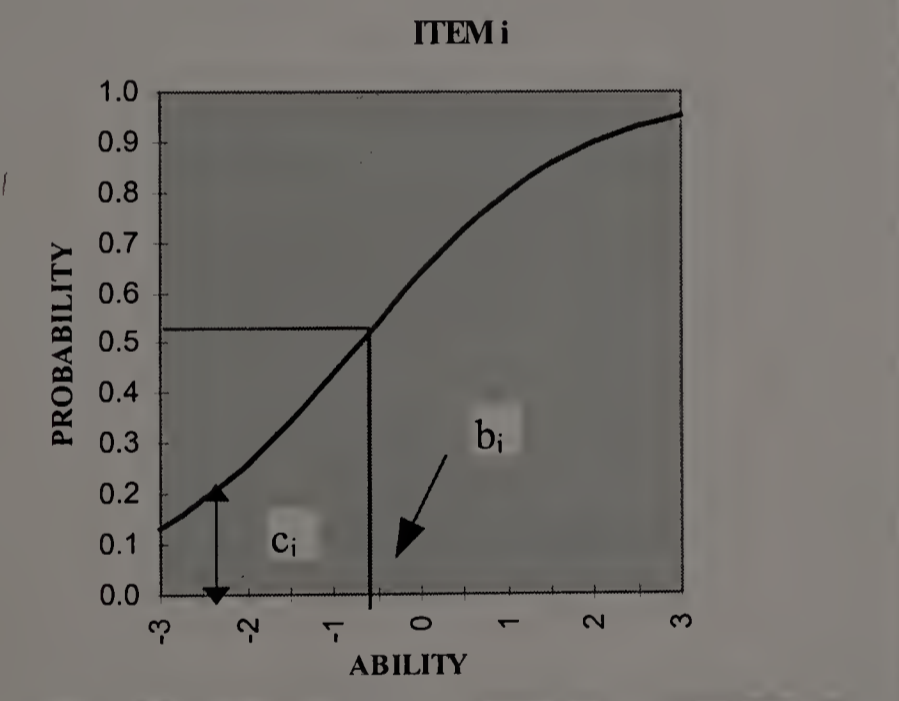
$D$  = scaling factor (adjustment to obtain logistic curve from the normal-ogive curve)

$b_i$  = item difficulty

$c_i$  = guessing parameter

$a_i$  = item discrimination

The concepts of item difficulty and item discrimination are important to understand while describing item response models. The ability of an item to discriminate among examinees at different levels of ability is called item discrimination. The item difficulty is the level of ability at which the item discriminates most effectively. These concepts become clearer when we visualize an item characteristic curve. An item characteristic curve is obtained when the probability function is plotted against the ability distribution of the examinees. A sample item characteristic curve follows:



**Figure 1: Typical Item Characteristic Curve**

The lower asymptote  $c$ , is the probability of a correct response by the lowest ability examinees, in other words, the guessing parameter. If items are being responded correctly by even the examinees with very low ability, the value of the lower asymptote

will be non-zero which corresponds to a non-zero guessing parameter. If  $c$  is zero the probability of a correct response corresponding to  $b$  on the ability scale is,  $(1+c)/2$ .

### 2.5.2 Two-Parameter Model

The point was mentioned earlier that the logistic curves and the normal-ogive curves are very similar in properties; Birnbaum (1968) introduced and later adjusted the two-parameter normal-ogive function by a scaling factor to obtain the two-parameter logistic function. The idea was to replace the normal-ogive model without having to change the interpretation of the parameters in the normal-ogive model. If the items differ in terms of their difficulty level as well as their discriminating power while the guessing is minimal, the two-parameter model best fits the data. In other words, the two-parameter model can be obtained if we omit the guessing factor from the three-parameter model. The exponential function of the two-parameter model is given as,

$$P_i(\theta) = \frac{\exp(Da_i(\theta - b_i))}{1 + \exp(Da_i(\theta - b_i))}$$

### 2.5.3 One-Parameter Model

The Rasch (1960) or one-parameter model requires one ability parameter ( $\theta$ ) for each person and one item difficulty parameter,  $b_i$ , for each item to represent the relationship between an examinee and a test item. In other words, it can be considered as a special case of the three-parameter logistic model where all items are assumed to have equal discriminating power and guessing is minimal (Hambleton & Swaminathan, 1985; Koch & Reckase, 1979). The one-parameter model is also used widely in adaptive testing due

to the ease of computation involved in this model. The exponential form of the 1-pl model is given as

$$P_i(\theta) = \frac{\exp(\theta - b_i)}{1 + \exp(\theta - b_i)}$$

The involvement of item response theory in the computer adaptive testing can be made clear if we highlight the components of a computer adaptive test.

## 2.6 Major Components of a Computer Adaptive Test

Carlson (1994) states,

The one-to-one interaction for individual testing no longer requires a test administrator who must be specifically trained and scheduled, and who frequently is quite expensive. A computer with an appropriate algorithm takes the test administrator's place.

The computer algorithm is the heart of computer adaptive testing which can only be understood if the various essential components of a CAT are highlighted. A detailed list of the various components of a computer adaptive test is compiled in the next section as characterized by Carlson (1994), Thissen (1990) and Hambleton et al. (1991):

### 2.6.1 Item Pool

A set of items, questions, or tasks is needed for a test. For a valid and reliable test, the quality of items is of special importance. It is important that the items are of a certain desired difficulty level, and they discriminate well among the examinee population taking the test. The items that have a very low or a very high difficulty level tend to have a low discriminating power. Also examinees tend to guess more on items with very high difficulty. While maintaining the balance between these characteristics, it is also essential to have items tailored to the content on which the ability is measured.

Thus a test specialist is faced with quite a few challenges while constructing an item pool. In adaptive testing, there is also a need of a large number of items in an item pool to have enough items to cover a wide range of abilities as well as to reduce the exposure to the same item repeatedly. The most widely used theory behind the computer adaptive testing, that is, Item Response Theory also requires a large sample of examinees to conduct item analyses and to test its various assumptions.

### **2.6.2 Item Response Model**

As mentioned in the above section, IRT is the most widely used theory behind computer adaptive testing. Adaptive testing can be thought of as one of the most successful applications of IRT (Kingsbury & Houser, 1993). An item response model describes the function through which we develop a relationship between getting a correct response on an item and the ability of a person, corrected for the various factors that can affect a response. This relationship then enables us to obtain an estimate of test taker's ability. The ability is either reported directly or converted into a score for ease of score interpretation. The choice of a model depends upon the kind of responses we expect on a test. A model that fits responses on items that do not involve guessing may not be appropriate for data in which a lot of guessing took place. Similarly, if items are targeted to discriminate among various test takers, the model should include an item discrimination factor. The selection of an item response model is hence critical for getting an accurate estimate of an individual's ability. Hambleton and Swaminathan (1985) summarized several models used currently. The list is as follows.



**Table 2.1: Item Response Models**

Type of responses expected	Model
Dichotomous	Latent Linear Perfect Scale Latent Distance One-, Two-, Three-Parameter Logistic/ Normal Four Parameter Logistic
Polytomous	Nominal Response Graded Response Partial Credit
Continuous	Continuous Response

The distinction among the models is obvious by their titles depending upon the type of responses on items on a test. The three-parameter logistic response model is the most commonly used model in a computer adaptive testing situation (Hambleton, Swaminathan & Rogers, 1991; Weiss, 1983). The main reason behind the three-parameter model being the most widely used model in CAT is that it generally fits the multiple choice data better than other models accounting for the fluctuations in item discrimination, item difficulty, and guessing factors (Hambleton, Swaminathan & Rogers, 1991). This reasoning seems very acceptable as the model accounts for three parameters that can affect an examinee score, instead of just one as in the case of Rasch model.

It is rare in case of multiple-choice tests to contain items that have equal discriminating power or that no guessing occurs during the test. Some researchers consider guessing as an “integral part” of adaptive testing and that any model which does not allow for guessing can provide misleading results (Wainer et al., 1990). Some modifications can be made to account for factors like guessing and discrimination if items are acceptably close to a certain model. For example, addition of many choices to items may reduce guessing or modifying the model slightly by keeping the discrimination parameter as a small constant value might improve model fit. Although three-parameter

model is generally considered best to fit the binary response data, the other models can be easy to work with and may involve lesser costs. The one- and two- parameter logistic and normal-ogive models are also used frequently due to such reasons (Carlson, 1994).

The item response models are considered fallible but in practice, it's difficult to declare a model to be appropriate for a set of responses (Traub, 1983). However effort should be made to select the best model by using several goodness of fit procedures and indices. One of the existing calibration procedures is selected to estimate the item parameters in order to test the model fit (comparing estimated with true parameters).

### **2.6.3 Starting Point and Initial Estimate of Ability**

Test length can be affected by the difficulty level of the initial item (Weiss & Kingsbury, 1984). The closer it is to the ability estimate, the lesser number of items are needed to reach to the final estimate of ability. The selection of the first item, however, is a complex decision to make. If the initial item is too easy, the test taker may take it too lightly and make careless mistakes. If the item is too hard, the test taker may become nervous at the very beginning of the test. The situation becomes worse if the examinee guesses the answer, since the next item will be more difficult. It is therefore desirable to have as much prior information about the examinee as possible in order to administer an item at the appropriate difficulty level. In some cases when there is no prior information available, the CAT algorithm starts by asking some background questions. Following is an example of questions that were chosen for a language proficiency test to select a starting point (Laurier, 1990):

- How many years did you study this language?
- Did you ever live in an environment where this language is spoken? If so how long ago?

- How do you rate your proficiency level on a scale of 1 (beginner) to 7 (advanced)?

The responses to such questions are then used to obtain a preliminary estimate of the examinee's ability. If there is not much background information available then item with the medium difficulty is chosen. From a psychometric point of view, adaptive tests will be more efficient if the initial item is of middle difficulty because it is the best estimate of an examinee's ability if no background information is provided (Mills & Stocking, 1996).

#### 2.6.4 Item Selection Strategy

A computer adaptive test is tailored to the examinee's level of ability. This means, the selection of an item depends upon the responses to the previous items. If the response to an item is correct, the next chosen item will be more difficult, while if it is wrong an easier item will be chosen and the ability is recalculated. In order to achieve this selection, several branching techniques can be used. However, IRT involves procedures that have resulted in a considerably efficient selection of items, maximum information selection and Bayesian item selection techniques being the most promising ones (Hambleton, Pieters & Zaal, 1991).

In maximum information selection technique, at a certain ability level, an item is chosen from the item pool that gives the maximum information about the examinee at that ability level. The "information" is provided by the *item information function* mathematically described as,

$$I(\theta) = \frac{[P'(\theta)]^2}{P(\theta)Q(\theta)}$$

where  $P'(\theta)$  is the first derivative of  $P(\theta)$  and  $Q(\theta)$  is the probability of getting an item wrong at an ability level  $\theta$ .

This item selection approach is designed to maximize measurement precision and has been preferred by many adaptive testing researchers (Kingsbury & Wiess, 1983) because it does not make prior judgments about the ability distribution. Although this is the most common technique used to select items, other item selection techniques have been considered in the past.

Bayesian item selection approach is another commonly used approach, which uses the posterior variance as the criterion of item selection. In other words, an item that minimizes the variance of the posterior ability estimate based on the responses to previous items is selected. Using a Bayesian technique solves some of the problems encountered with the maximum likelihood approach (Swaminathan & Gifford, 1982; Swaminathan & Gifford, 1983). Bayesian estimates can be obtained for zero items correct and for perfect as well as aberrant response patterns (Swaminathan & Gifford, 1985). A number of simulation studies have shown that Bayesian adaptive testing technique results in stable, reliable, and valid scores even for very short tests (McBride & Wiess, 1983; Jensema, 1974). Researchers continue to find better techniques, such as finding utility functions for an examinee to maximize score performance when selecting items or using global information to provide an ability estimate closer to the true ability estimate (Chang, 1996; Chang & Ying, 1997).

### 2.6.5 Computation of the Provisional Estimate of Ability

A number of methods currently exist to estimate the ability and item parameters.

Following is the list of some of the commonly used methods that can be used for parameter estimation (Hambleton, Swaminathan & Rogers, 1991).

**Table 2.2: Ability and Item Parameter Estimation Procedures**

Estimation Procedure	Model (Brief Explanation)
Joint maximum likelihood (Lord, 1974, 1980)	One-, Two-, Three- Parameter model (The ability and item parameters are estimated simultaneously)
Marginal maximum likelihood (Bock & Aitkin, 1981)	One-, Two-, Three- Parameter model (The item parameters are estimated with ability parameters integrated out)
Conditional maximum likelihood (Anderson, 1972, 1973; Rasch, 1960)	One-Parameter model only (The likelihood function is conditioned on the number right score and the item parameters are estimated)
Joint & Marginal Bayesian estimation (Mislevy, 1986; Swaminathan & Gifford, 1982)	One-, Two-, Three- Parameter model (The ability is estimated with joint/marginal estimation of item parameters)
Heuristic estimation procedure (Urry, 1974, 1978)	Two-, Three- Parameter model

Currently, the two most commonly used estimation procedures for dichotomous responses are maximum likelihood, marginal maximum likelihood and Bayesian estimation (Hambleton, Swaminathan & Rogers, 1991). Maximum Likelihood estimation involves finding the maximum of the logarithm of the joint likelihood function (L) which is the product of the probabilities ( $P_j$ ) of correct response at a given ability level. The Maximum Likelihood estimate of ability is then the examinee's ability, which makes the likelihood function, a maximum. This relationship can be given as,

$$L(u_1, u_2, \dots, u_n | \theta) = \prod_{j=1}^n P_j^u Q_j^{(1-u)}$$

This method can be very useful, however, the method fails when the number of items is small and there is less chance of mixed responses. The Bayesian estimation is a possible solution to this problem as it's not dependent on the number of correct or incorrect responses. This technique basically involves modifying the likelihood function to include any prior information obtained on the ability distribution (Swaminathan, 1983). The relationship can be described as,

$$f(\theta|u) = L(u|\theta)f(\theta)$$

Here,  $f(\theta|u)$  is the posterior density, where the mode of this function is the “most likely to be” value of  $\theta$  and can be considered as our desired ability estimate,  $L(u|\theta)$  is the likelihood function, and  $f(\theta)$  gives us the prior information on the distribution of  $\theta$ . It avoids the problem of response pattern dependency; however, the estimates may be biased (Hambleton, Pieters & Zaal, 1991).

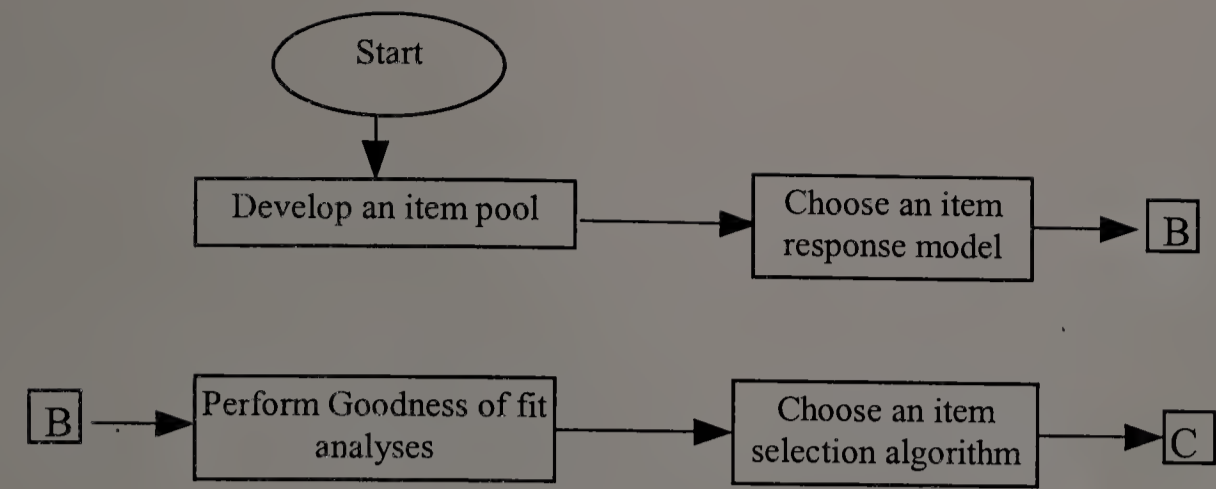
### 2.6.6 Termination Criterion

An important consideration in the adaptive testing is when to stop administering the test. In most of the cases, measurement precision or test length is the basis for terminating a CAT algorithm, however, time may also be a consideration. The CAT algorithm can continue until some desired value of standard error of measurement is achieved thus varying the number of items administered to each examinee. It is also possible to specify different levels of measurement error criteria depending on the ability range; more care should be taken at the middle level abilities. It is also possible to fix the

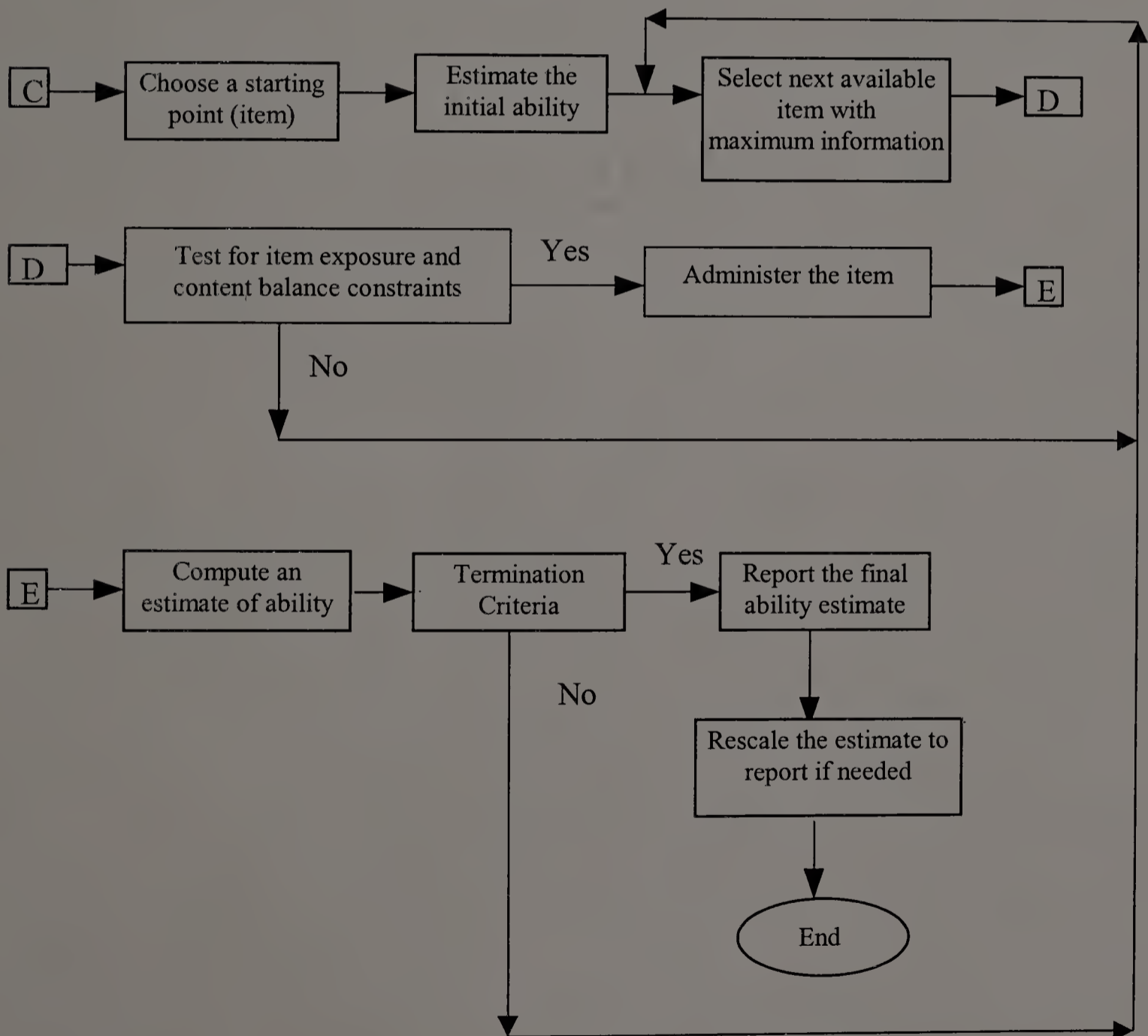
testing time and hence get ability estimates with different errors of measurement. It is also possible to combine the two conditions, that is, have a fixed minimum number of items and a desired measurement error. Numerous studies have been conducted to analyze various stopping rules. Stocking (1987), for example, discovered that by using standard error of measurement as the stopping rule, it was sufficient to know the examinee's true score and the number of items administered to predict whether the true score is over or under estimated. Bergstrom & Gershon (1992) on the other hand used a stopping rule based on the confidence in the pass/fail decision in a medical exam where items were targeted to the ability of the examinee. The researchers suggested that using such a stopping rule in a computer algorithm that uses maximum likelihood estimation and Rasch model for item calibration is most efficient.

#### **2.6.7 Method for Computing the Final Estimate of Ability**

It is usually the same as step 2.6.5; however, another technique can be used to estimate the final ability. The following flow chart illustrates the utility of various components of a CAT during an administration algorithm:



**TEST ADMINISTRATION FOR AN EXAMINEE**



**Figure 2: Flowchart for a Computer Adaptive Test Administration**



## CHAPTER 3

### CONSIDERATIONS IN THE DEVELOPMENT OF A CAT

#### 3.1 Brief Overview

The occurrence of response aberrance in a CAT cannot be understood without a clear description of the related issues in a CAT design. The purpose of this part of the review is to describe some important considerations in CAT development process. The chapter also highlights some of the assumptions that are made for our case study to look at the aberrant response patterns in a CAT. The last few sections focus on the examinee's interaction with a CAT leading on to the discussion of response pattern aberrance.

##### 3.1.1 Development of an Item Pool

Since the estimate of ability and the choice of the next item administered requires the knowledge of the item parameters, one of the main requirements to implement a computer adaptive test is a calibrated item pool. A brief explanation of the importance of an item pool was given in the section describing the various components, however, a detailed description of a calibrated item pool or item bank is needed at this point.

The introduction of item response theory and the advent of computers have made the creation of item pools a reality (van der Linden, 1986). A set of test items, all measuring the abilities of an examinee population on a similar construct domain can be considered as the beginning of an item pool. At this point all we know about the pool is some initial estimates of the quality of items from our previous knowledge about the

nature of the test. As soon as a test is administered from the pool, a set of responses is available. These responses can be used to score the test as well as to find the parameter estimates based on a model selected on the basis of assumptions made about the nature of items. The process of estimating item parameters can be viewed as placing the items on a measurement scale, commonly known as item calibration. In the case when the parameter estimates are obtained for new items to be included in subsequent item pool from the responses of an examinee during a testing session, it is called on-line calibration (van der Linden, 1986), currently used in many adaptive tests. As soon as the item parameters are known, the ability parameters can be estimated using the model as the measurement model.

An ideal item pool consists of sufficient number of items whose measure of precision follows a rectangular distribution across the entire ability range to be measured. However, it is a challenge to achieve such an ideal; in practice there are many constraints that come into play while deciding on an adequate item pool in terms of structure and size of the pool as well as the quality of items e.g. content representation, item overlap etc.

### **3.1.2 Item Exposure**

In conventional testing programs, a set of questions is administered to a large group of examinees on a single day. Thus item exposure is limited to a short period of time. In adaptive testing, however, the period of time in which items are exposed is increased, therefore, a problem of a higher item exposure rate arises, especially for more popular items. Most of the testing agencies incorporate item exposure control mechanisms into the adaptive testing algorithm. A decision is made before the test

administration, about the rate at which the items will be exposed to examinees. For example, no more than 20% of the examinee population will see a given item in a large scale test (Stocking, 1996; Schaeffer et al., 1995) or 7% in case of smaller scale tests (Leucht, et al., 1996). A question arises at this point; what is the most effective way to reduce item exposure? Is it a good idea to add more items to a pool to reduce exposure within the pool or should there be multiple pools with constant exposure rates? Item exposure, in addition to other factors, largely depends upon the size and depth of the item pool (Simpson & Hetter, 1985).

The Simpson and Hetter methodology employs using a large pool of items (both discrete and within sets) and development of exposure control parameters for individual elements in the pool. The exposure control parameters are lowered for items that might be frequently chosen based on their content and/or statistical properties. This implies that these items are only administered for some fraction of times they are selected. Similarly these parameters are raised for items that are less desirable to be chosen. A well-known fact is that using large item pools eliminates many of the security risks and also reduces the negative effects of implementing item exposure controls (Leucht et al., 1996; Mills & Stocking, 1996).

In addition to Simpson and Hetter methodology, a number of techniques have been recently proposed for reducing exposure rates. A brief definition of some of those methodologies is given below:

- a) Simpson & Hetter Conditional Methodology: The technique is an extension of the S & H methodology that was described in the previous section. This method limits the item exposure differentially across ability levels. In other

words, the exposure is controlled conditioned on ability (Stocking & Lewis, 1995)

- b) A-Stratified: In this approach, the item pool is divided into strata based upon the level of item discrimination parameters. The strata with higher a-values are used as the test progresses; within each stratum, the item that has the b-value closest to the examinee's current ability estimated is chosen (Chang & Ying, 1999).
- c) A-Stratified with b-blocking: This method extends a-stratified method by forcing b-parameter values to be evenly distributed across all strata. Both a- and b-parameters are, therefore, used in forming the strata (Chang et al., 2000)
- d) Tri-Conditional: This method combines Conditional Simpson & Hetter methodology with additional conditioning on context. The technique, in other words, conditions the exposure control on item, ability and test context (Parshall et al., 2000).
- e) Stochastic: This method of exposure of control will be employed in this study due to its simplified implementation. The method was proposed by Reveulta and Ponsoda (1998) and Robin (1999). In this method, items are not allowed to be administered more than  $100k\%$  of the tests where k is the maximum exposure rate. Suppose that a test is administered t times and a is the number of times a particular item has been administered in the previous t tests. The exposure rate of that item will be  $a/t$  that has to be less than k for the item to be administered. The item will be available for some tests, and then will be unavailable for some other tests. The item will be available again as the

quotient  $a/t$  decreases and becomes less than  $k$ . For example, if the value of  $k$  is 0.25,  $t$  is 50 and  $a$  for a particular item is 15, the exposure rate of that item is hence 0.3. This item will be unavailable for administration until the exposure rate becomes less than 0.24.

### 3.1.3 Constrained Item Selection

In a computerized adaptive test, skilled test developers make up an item pool that satisfies the usual requirements of matching content specifications. The content validity is particularly important in the case of achievement testing where an individual's achievement is measured in a number of content areas at the same time. However the issue of content balancing becomes problematic in a CAT. In the adaptive testing, since the examinees are presented only with items suited to their ability, it is essential to have items representing each content area covering a range of difficulty levels (Schartz, 1986). This leads to a large number of items in an item pool representing a wide range of ability for various content areas. In addition test developers are faced with the constraints of preventing item overlap i.e. item giving away the answer to another item, and item block maintenance i.e. keeping item in its current block. In summary, for item selection, the algorithm must satisfy constraints on content as well as statistical and intrinsic properties of items (Eignor, 1993).

Hence, in the case of adaptive testing, there may be some trade-off in terms of determining which specifications are important and which can be relaxed to take full advantage of the measurement models as well as other adaptive test requirements.

Stocking and Mills (1996) emphasized the fact that a careful examination needs to be done of the content specifications as well as other constraints by relating them to the construct being measured. The purpose is to insure whether or not some constraints could be relaxed. It is feasible to apply different weights according to the importance of certain specifications and restrictions. The adaptive testing algorithm must maintain a record of the extent to which the test meets each condition and select items in a way that best fulfills the specifications. Item selection strategies such as optimal constrained adaptive testing where all items are selected in the form of a "shadow test" at the current ability estimate (van der Linden & Reese, 1998) or selection with the sequential probability ratio test (Eggen, 1999) are being introduced. However, the **Weighted Deviations Model** using Fisher's information function still remains popular. The adaptive test algorithm implied in this study used the same Weighted Deviations Model (WDM) for item selection (Stocking & Swanson, 1993, Stocking, 1996).

In the WDM approach to item selection, the next item that is selected for administration is the item that simultaneously

- (1) is as informative as possible at a test taker's ability level
- (2) contributes as much as possible to the satisfaction of all other constraints in addition to constraints on the item information

At the same time, it is required that the item

- (3) does not appear in an overlap group containing an item already administered
- (4) is in the current block as the previous item or starts a new block

### 3.1.4 Dimensionality

Most item response theory models assume that a single ability or latent trait can explain an examinee's test performance. However, it is not a simple matter to construct items which all measure the same trait, there are many cognitive factors that may account for an individual's response to an item (Traub, 1983; Keitzberg, Stocking & Swanson, 1978; Ackerman, 1987). For a group of individuals, it is doubtful that each person to respond to a single item would use a single cognitive skill or a constant combination of skills. For example, a licensure exam may measure several skills, or a history test may measure a composite skill of history knowledge, reading and memorization. Wang, Wilson and Adams (1995) call these types of multidimensional situations as between-item and within-item respectively. The assumption of unidimensionality has always undergone a lot of criticism. Laurier (1990) while describing the concerns in the application of CAT to various types of testing considers the assumption of unidimensionality as the most "formidable" problem. According to Laurier, CAT cannot be applied on a Cloze test (where finding a correct word in a context increases the chance of finding the next word) because it doesn't satisfy the condition of independence which is a type of unidimensionality. He states

It (CAT) should never be used to create a diagnostic test that aims at finding weaknesses or strengths on various discrete points because this type of test is not unidimensional. By the same token, it should not be used on so-called "communicative" tests that attempt to measure aspects of the communicative competence without isolating the different dimensions in separate sub-tests.

Traub and Wolf (1981) expressed their difficulty in understanding how a CAT could serve useful achievement and diagnostic purposes if constructs like analytical

reasoning and reading comprehension are assumed to be unidimensional and assumed not to be influenced by factors like instruction, type of text, language background, etc.

Research has shown that no matter how much effort is put into construction of items to maintain unidimensionality of an item pool, it is almost impossible to achieve this aim for every individual. Wainer (1983) found that the trait being measured, although unidimensional for most of the test population, may be multidimensional for some small sub-population. Research, however, also indicated that although the most commonly used theory behind CAT relies on unidimensionality, empirical results show that the model is suitable when the items in the pool have one dominant dimension (Green et al., 1984; Drasgow & Parsons, 1983). Items related to small secondary dimensions would tend to have smaller item discrimination values but will not affect ability estimate a lot. Thus, a necessary requirement for CAT is that either the item pool is unidimensional or has one dominant dimension. The implication in this study of an item pool is that the pool is *unidimensional*. However the issue of response aberrance tends to violate this assumption in some cases; this will be discussed later in the chapter.

## **3.2 Examinee Interaction and Test Taking Behaviors within a CAT Environment**

### **3.2.1 Examinee Interaction with a CAT**

Although it seems that the introduction of computers to administer test makes things easier on the part of examinees, it's not quite what we observe. The experience of taking a CAT is not as simple as taking a conventional paper and pencil test. Apparently it's just a question appearing on a computer screen and examinee choosing an answer till



the test terminates. However, there are a number of unique aspects to CAT that might influence an examinee's performance.

First, computer based testing is not a familiar mode of testing to many test takers. Some people might be irritated by looking at the bright computer screen while others may find it fatiguing to scroll through long items. In reality, many people tend not to talk about such aspects of a CAT to hide their unease with the computers in general, not to look "computer illiterate" (personal experience). In reality, many people even feel uneasy with answering through a keyboard or a mouse (Wise, 1997). Another factor that might cause anxiety among test takers is the feeling of items getting easier or harder based upon their responses. In a CAT, easier items mean poor performance resulting in anxiety in a test taker. Also, most of the examinees have some idea that there is a computer algorithm, which selects the next item, presented to them and that the CAT is shorter and more precise. This creates an additional pressure on a test taker as he or she knows that each item has a larger impact on their final score (Wise, 1997).

Another aspect of CAT that examinees have frequently reported as discomforting is the inability to browse, skip through and go back to their answers (Vispoel et al, 1994; Wise, 1996). Wise (1996) argued that the examinees are likely to gain scores if they were allowed to review and rethink their answer. Denying item review may result in increased levels of anxiety as it results in a lack of control over the test.

Testing time limits also have a great impact on the test takers. Placing a time limit on a test is a typical feature of a standardized test. However, time limits serve only the interests of the test administrators and work negatively for the test takers. It is even more complicated when it comes to CAT, whether it's fixed or variable length. In case of

fixed length CATs, if able examinees are getting harder items and less able examinees are getting easier items, the same time limit does not make much sense, as harder items tend to require more time to answer. CATs that use measurement precision of an ability estimate as the stopping-rule end up in different test lengths for different examinees. It is therefore hard to decide a final limit on a testing time.

The above mentioned factors might result in test anxiety for examinees and test anxiety has proven to affect examinee-performance adversely (Hills, 1984; Wise, 1996). Another related facet of a CAT that has been researched extensively is the issue of response times when a time limit is in place for a test. The following section discusses the issue of people responding differently in terms of the time that they take to complete a test.

### **3.2.2 Examinee Response Times and Test Taking Behavior**

Examinees with the same ability require different amounts of time to complete a test item. Researchers have found that many times differences in test-performance may be due in part to the differences in the time people take to respond to questions. In other words, the difference in response times might affect people's performance instead of their knowledge, skills or ability. It is therefore important for test takers to pace themselves while taking a test.

As mentioned in the previous section, it is possible for test takers to get anxious thus taking long to respond or they may just be slow thinkers. Differences in response-times of items administered early in the test may not affect performance on those items, but may affect the performance towards the later part of the test. As the test taker moves

to the later section of the test, either he or she has to leave the items as unanswered or simply guess randomly. Schnipke & Scrams (1995) refer to such guessing behavior as the “rapid-guessing behavior”. The researchers indicate that it is easy to track such behaviors as some test takers respond so quickly towards the end of the test that one could not even read the item in that time. Such factors cause the test to be speeded for some examinees than the others. Response-time analyses have shown, for example, that Hispanics and African Americans spend more time on each item than other groups (O’Neil & Powers, 1993; Llabre & Froman, 1987). Also, some groups are better in allocating their time for an item according to the item difficulty (Schaffer et al, 1993).

Schnipke and Pashley (1997) conducted a research study on the response timings of two groups of native and non-native English speakers for nationally administered high-stakes reasoning test. The test was a 25-item computer-based but non-adaptive test. The researchers used survival analytic techniques to look at the distribution of response times. The researchers found that the test score was a significant predictor of response-time for all items. On the first half of the test, non-native English speakers responded slower on average than native speakers while on the last half, they responded much faster and rushed by guessing randomly. Generally studies have found that increased overall testing time improves scores, but no significant interaction was found with race or gender (Wild et al., 1982). Another related finding that was made clear through this study was that computer based testing made it much easier to keep track of the response times by each examinee for each item. Examinee’s average response time could be obtained by combining examinee’s response times on all items. Similarly item’s average response time could be obtained by combining all examinees’ responses on that item.

Although computer based testing makes it easier to keep track of the response times; the issue is not so simple when it comes to the computer adaptive testing. Since, each test in a CAT has a unique set of items; the expected completion time is different for each test and thus each examinee. For a variable length CAT, the expected completion time does not only depend on what items are used but also on how many items are used. Hence, it is difficult for test administrators to decide on appropriate time limits on a CAT and it is even more difficult for examinees since, unlike P&P tests, they are not exposed to the whole test to pace themselves accordingly. In addition, the IRT assumption of unidimensionality also gets violated if the extraneous variable of timing impacts the performance of some examinees in addition to the construct being measured (Bontempo & Julian, 1997; Oshima, 1994). If we look at the past research, we observe that there are many unanswered questions when it comes to response-times and pacing correctly on a CAT.

Regression studies on response-times on a CAT fail to explain the variation in response-times. For example, Kingsbury et al (1993) found that none of the variables that they included in their study of response-times on a CAT accounted for more than 8% of the variance in response-times. The studies where the variance did get explained, the results were not what would be expected.

Bergstrom and Gershon (1994) analyzed response-times on a computer adaptive test using a hierarchical linear model. Their finding was in contradiction to some of the earlier findings. The researchers included several within and between persons variables in their study. The within-variables included difficulty, position, length, graphical nature, content category and position of answer key for an item. The between-variables included

test-anxiety, ethnic background, gender, language, age, and final ability estimate of the examinees. Although there were significant differences between examinees, much more variation was found within an examinee. The researchers found that the examinees spent more time on items they got wrong than on items they got right. Factors like position of an answer key, item length, relative item difficulty and the test anxiety significantly affected response times. Although, for some examinees it was clear what affected their response times but for others, the researchers indicated little understanding of why they spent more time on some items than the others.

Van der Linden et al. (1999) conducted a study to look at the response-time distributions in a simulated CAT for ASVAB data. The researchers used a statistical model for examinee response-time distributions. The predictions from the model were used as constraints on further item selection to adjust for the speededness for all examinees. The researchers observed that the algorithm reduced the effects of speededness for examinees that would have otherwise suffered from the time limits. Interestingly all those examinees were high ability examinees. The study, which proved to be a positive step towards dealing with response-time issue, also re-enforced something that was found in the Bergstrom study, that is, ability and response-time seem to be uncorrelated on a CAT.

The same finding was observed when Swanson et al. (1997) conducted studies on response times for the National Board of Examiners' Step 1 and 2 Licensure exams. It was found that the ability and response-times were uncorrelated generally, however a moderate positive correlation was observed if the time limit was too restricted. These studies while looking at ways to handle response-time issues to adjust for speededness

also suggest that in a CAT, unlike conventional tests, it's not only the low ability candidates who get affected by restricted time limits. Researchers are looking at the issue of response-times on CAT very carefully, however very few recommendations have been made to date.

Steffen and Way (1999) evaluated the different strategies that examinees might want to take while taking a CAT in terms of the time spent on items. The researchers found that the scores for high ability examinees would be negatively affected if those examinees went slowly on the early items their scores, as they would end up guessing on the items that they actually knew. Low ability examinees, on the other hand could do the best if they spent more time on the early part of the test and guess towards the end. The middle ability examinees will suffer in term of their scores if they provided incorrect answers towards the early sections.

This leads to our issue of interest, that is, the issue of response aberrance for some examinees in a CAT. Although, there might be ways to go around the issue of taking the test efficiently in time, it does create unexpected patterns of responses. Spending more time on an early part of the test and rushing through the last part by guessing creates response patterns that do not fit to the response models working behind the CAT algorithm. Such response patterns are called "aberrant" and force the CAT algorithm to produce incorrect estimates of ability. The details of such scenarios are presented in the next section.

### **3.3 Aberrant Response Patterns**

#### **3.3.1 Definitions of an Aberrant Response in IRT Framework**

##### **3.3.1.1 Statistical Definition of an Aberrant Response in IRT Framework**

The lower the probability of the response determined by the IRT model parameters, the more aberrant the response (Reise & Due, 1991). In general, the aberrance is statistically defined in terms of the maximum likelihood function as the value of the function decreases due to the occurrence of an unlikely response, given the model. For example, a pattern of correct guessing on a set of difficult items by a low ability examinee will adversely affect the likelihood function, which in turn results in an inaccurate estimate of the final ability of the examinee.

##### **3.3.1.2 Definition of Aberrance in Terms of Information**

An aberrant response is the one that provides less psychometric information (Lord, 1980) for estimating ability than would be expected by the parameters of a specified IRT model. Here the aberrance is defined in terms of the test information function, as its value decreases if aberrant responses occur.

#### **3.3.2 Appropriate Measurement or Person-fit Research**

Since the occurrence of aberrant or non-model fitting responses for examinees frequently results in incorrect score reporting, the whole purpose of a test is hence defied. Over the past 25 years, this area received a lot of attention where researchers have tried to detect examinees with such non-model fitting response patterns or in other words “misfitting persons” or “inappropriate” score or response patterns. As mentioned before,

this area of research has been known as appropriateness measurement in the past (Yi & Neiring, 1999; Drasgow & Levine, 1986) and as person-fit more recently (Meijer & Neiring, 1995; Reise & Due, 1991). While almost all of the studies in this area have been conducted to address the issues of detection, they do provide us with an idea of the kind of response aberrancies that could be expected and specifically the kinds of simulation studies that would be suitable to various situations. Readers that are interested in a detailed overview of and recent developments in the person-fit detection methods, refer to Meijer and Sijtsma (1994).

### **3.3.2.1 Misfitting or Aberrant Reponse Patterns**

An item response model can be inappropriate for an examinee even though the model may be appropriate for the whole group of examinees. The model may be inappropriate or the responses may be aberrant for a number of reasons. Researchers have observed that in a paper and pencil testing situation, for example, examinees may skip an answer on the test without skipping the item on the answer sheet. In some cases, they might turn easy items into "tricky" hard questions thus creating difficulty in the items that was not in the test design. (Mcleod & Lewis, 1996). For the remainder of the test, the IRT model falsely assumes that the examinees are answering the items based on their true abilities thus resulting in low scores for such examinees.

Another situation arises, when examinees cheat or copy some answers from the other test takers. The ability estimate in this case will depend on the other test takers' abilities and may result in unexpectedly high scores if the other test taker is a high ability examinee.



The inappropriateness of a response according to a model, might also be due to the violations of the underlying assumptions such as invariant ability over items/subtests, unidimensionality or local independence assumption in case of most commonly used IRT models (Glas & Meijer, 1998). Guessing, cheating, memorization, creativity, fumbling, and fatigue, for example, result in the violation of the assumption of invariant ability across items. Cheating and memorization also results in the violation of the assumption of unidimensionality.

The issues are more serious in case of computer adaptive tests that bring along with them numerous allowances but also constraints such as the prohibition of item review or item omits. The examinees therefore, intentionally or unintentionally come up with innovative techniques to beat the test. In a CAT, for example, in addition to the above-mentioned behaviors, the issue of memorization can be more serious compared to paper and pencil tests. Research shows that if the item pool is smaller, the examinees might inflate their scores by memorizing difficult items and thus routing themselves to more memorized items (McLeod & Lewis, 1996).

While a number of aberrant behaviors may occur in a CAT, typical forms of aberrant response behavior are guessing and cheating which may result in spuriously high or low scores (Glas & Meijer, 1998). There has been little research on the ways to detect aberrance in a CAT, assuming that the same indices could be used as in conventional test. However, researchers like Bradlow et al. (1998) and Glas & Meijer (1998) have shown that since, in a CAT, examinees receive different items in varying orders and there are no missing data (because of no-skipping constraint), traditional indices may have lower

power. An explanation of the recommended indices and a comparison of such indices with conventional indices is outside the scope of this study.

### 3.3.2.2 Simulation Studies

An important aspect of person-fit research is to simulate particular response patterns for different groups of examinees. A wide array of literature exists where researchers simulated a variety of response patterns for exploring detection indices. However, very few studies were conducted to represent response aberrance in a computer adaptive testing situation. The following section briefly illustrates two such studies providing us with an idea of the nature of such simulations.

McLeod and Lewis (1996) conducted a comprehensive simulation study to compare two person-fit indices to detect memorizers in a CAT administration. Five levels of ability were chosen (-1.0,-0.5,0,0.5,1.0). Three sets of ten item difficulties were selected at each ability level; the difficulties were chosen to reflect the levels that might arise in adaptive tests for examinees at the five ability levels, and ranged from -2.0 to 1.0. Five different response patterns were then simulated for each item set, based on the ability level associated with that test.

For the first response pattern, a Guttman pattern was used, that is, the simulees gave correct responses to all easy items and incorrect responses to all the items with difficulties above their abilities. This represented a perfectly appropriate response pattern. The second pattern simulated the Reverse-Guttman pattern where simulees provided correct responses to all difficult and incorrect responses to all easy items. This represented the case where low ability examinees had memorized the difficult items. This

pattern was then manipulated to depict another pattern, where the examinee correctly answers few easy and incorrectly answers few difficult items. The Manipulated Reverse-Guttman was meant to fit the IRT model better than the Reverse-Guttman but not as well as Guttman. The fourth pattern was generated according to the normal IRT model while the fifth pattern simulated random responses. The person-fit index was therefore expected to detect few of the normal and Guttman patterns as non-fitting, consistently detect Reverse-Guttman and have varying levels of detection for random and manipulated Reverse-Guttman response patterns.

The second part of the study involved data from actual CAT administration of GRE where 1650 response patterns were generated. The fifty most-frequently exposed items for the top 5% of the examinees were considered to be memorized. Results of the study are behind the scope of this study.

Another interesting simulation study was performed by Glas and Meijer (1998). The study is a replication of the earlier study by Klaur (1995) extending the use of the proposed indices to the CAT environment. The researchers simulated data to depict response patterns that violate two IRT assumptions for a 2-PL model; invariant ability across subtests, local independence between items. The data were simulated to depict aberrant responses by defining alternative models. The non-invariant abilities across subtests were modeled by assuming that the 2-PL is valid during the whole testing session but that the respondent's ability parameter changes during test-taking. It also assumed that a person has two ability parameters; the first parameter governs the responses on first half of the test while the second governs the second part. The lack of local independence was simulated by assuming that the probability of a correct response on an item is

augmented by a previous correct response. This was obtained by introducing a transfer parameter. Again, the results of the study are behind the scope of this paper and are not included here.

The above mentioned simulation studies are excellent examples of simulation studies that could be conducted in a CAT environment to depict aberrant response behaviors

### **3.3.3 Occurrence of Aberrant Response Patterns in a CAT**

#### **3.3.3.1 Scenario 1**

The first factor is related to examinees pacing on a test. As mentioned in the last chapter, the concept of Pacing is also referred to as an examinee's "time management" for completing the test. For example examinees might spend an inordinate amount of time in correctly answering a certain percentage of items on the test in the beginning and spend a very short time towards the end. This might result in examinees guessing extensively towards the end thus resulting in inaccurate provisional estimates of ability. The inaccurate pattern towards the end lowers the final ability estimate compared to the item difficulty while the test was well targeted in the beginning. This sort of behavior has frequently been observed in computer adaptive testing situations. For example, Bontempo & Julian (1997) found that in NCLEX (1996 administration), 77% of the examinees (out of those examinees whose tests consisted of 215 of a maximum of 265 items and took more than 4 hour; 10 minutes out of 4 hour; 45 minutes) rushed and guessed rapidly towards the end.

### **3.3.3.2 Scenario 2**

The examinees might answer hard items correctly and easy items incorrectly. This situation arises when a student studies intensively to answer difficult questions and in that effort ignores simple and easy questions. Here, the assumption is that difficult items correspond to difficult subject matter. The examinee's response pattern will deviate from the rest of the group with similar abilities, given the most common probabilistic models (Meijer & Sijtsma, 1994). A variation on this situation might be that a candidate cheats to get difficult questions correct on a certain portion of a test (see section 3.3.2).

### **3.3.3.3 Scenario 3**

Aberrant response patterns are likely to occur in a situation where the distractors for an item could be partially correct. However, the examinee can detect such distractors only if they are exceptionally creative. High ability examinees can therefore get such items wrong against the model expectation. Also in this case, the construct of creativity interferes in the item selection algorithm as the underlying assumption of unidimensionality no longer holds.

### **3.3.3.4 Scenario 4**

Item selection algorithm is less responsive to the candidate's responses. This might occur due to the shortage of difficult items in the item pool. Hence, the provisional ability estimates are constantly higher than the item difficulty and the test ends before examinees could get harder questions. On the other hand, the estimates could be consistently lower than the difficulty of the items delivered if there are fewer easy items

on the test. It also increases the test anxiety for those examinees. A number of item selection constraints such as content constraints and item exposure controls could also limit the selection of items targeted to the ability estimates.

The first three situations relate directly to aberrance as they are known to actually cause aberrance. The last scenario, however, interacts with aberrance in that it can be caused by aberrance due to unexpected and abnormal pool utilization.

## CHAPTER 4

### METHODOLOGY

#### 4.1 Introduction

The focus of this study was to identify the conditions under which aberrant response patterns become particularly problematic in a computer adaptive testing environment. The study examined the effect of test length and time limits on the occurrence of aberrant response patterns when examinees rush into random guessing towards the end of a CAT.

In this study, four different testing scenarios were examined; fixed length performance tests with and without content constraints, fixed length mastery tests and variable length mastery tests without content constraints. For each of these testing scenarios, the effect of two test lengths, five different timing conditions and the interaction between these factors with three ability levels on ability estimation were examined. For performance tests, the lack of items in a certain content area was simulated to look at the effect of their interaction with aberrance. For fixed and variable length mastery tests, decision accuracy was also looked at in addition to the estimation accuracy.

The interaction of aberrance with the total pool information was also studied briefly during the course of the research; however, the results of that particular analysis are not central to the study.

The time limits were imposed by the introduction of random guessing after the examinee had answered a certain percentage of items out of the total test length. The

response patterns were simulated first by simulating the item and ability parameters and then using the item parameters obtained from a high stakes test results.

Various indicators including Root Mean Square Error (RMSE) index, Bias index, and test information functions judged the effect of aberrance on the ability estimation. The distributions of false hits (incorrect pass/fail decision) were used to look at the decision accuracy in classification testing.

#### **4.2 Data Generation Model**

Data can be generated in a number of ways to mimic aberrant response patterns. For example, data can be generated conforming to an IRT model and then statistically manipulated to simulate different aberrancies in the response patterns. For example, changing certain percentage of correct responses to incorrect responses and vice-versa. Another method is based on the test information approach, where a model is conceptualized to generate data that simulate examinee response aberrance. Such models are commonly referred to as the Generalized Response Aberrance Models. Strandmark & Linn (1987) introduced one such model where aberrance was simulated by manipulating the item discrimination parameter for some individuals. This in turn manipulated the information provided by the underlying IRT model as the test information depends largely on the a-parameter. Such data generation models have been used in a variety of research studies, however, these techniques produce artificial response patterns instead of response patterns that may occur in real situations (Yi & Neiring, 1999). In our study, a data generation technique was used that reflected the examinees' behaviors more accurately.



### 4.3 Design

In the first step of the study, an item pool of 600 items was established using simulated item parameters. A fully adaptive test was administered to 1200 simulees (100 examinees at 12 ability levels) on 30 and 75 item fixed length tests using CBTS (Robin, 2000). For each test, every item was taken as a multiple-choice item with 4 alternatives. The CAT program used Weighted Deviations Model for item selection and Stochastic Item Exposure control methodology to control for item exposure (for details, see section 2 of the literature review). This administration of the CAT was called the CAT delivered under “Null” conditions. The “experimental” conditions included simulations conducted to generate data depicting response aberrance. To simulate random guessing at certain points in the test, the probability of a correct response was changed to the chance probability. Such conceptual framework for simulating random guessing behavior has been frequently used for detecting such aberrance using IRT based person-fit indices (Kogut, 1986; Meijer, 1994). The chance probability in this case was 0.25 because of the four alternatives to each item. In the experimental conditions, guessing was introduced after a certain percentage of items had been delivered. The following formula was used:

$$\text{Number of items (n)} = \text{factor} \times \text{total number of items}/100 \quad (1)$$

where factor is the percentage of items. Hence the regular administration of the CAT continued till the algorithm hit equation 1. At this time, the probability of the correct response was changed to:

$$P(\theta) = 1/\text{number of alternatives} \quad (2)$$

The following logic applies: For a fully computer adaptive test, an item is selected based on the information that items in the pool carry at the provisional ability estimate

and administered after other constraints have been applied. The ability estimate is computed using Bayesian estimation (see chapter 2 for details) based upon the responses to the previously delivered items. If responses to the first  $n$  items are generated according to the item response model (3PL in our case), the random guess at the  $(n+1)^{\text{th}}$  item disrupts the selection algorithm. The ability estimate that was expected for that particular examinee on an item with a particular item difficulty will not be produced. In fact, a wrong provisional ability estimate will be computed unless the guessed answer is the same as the expected answer. The information function will then be calculated for the remaining items in the pool. Hence the  $(n+2)^{\text{th}}$  item will be the one that provides maximum information at an incorrect ability estimate, thus resulting in an item which is not well targeted to the person's true ability. Since, we assume that the random guessing will continue till the end, the same process will be repeated again and again thus resulting in a final ability estimate much different from the true estimate.

The experimental conditions were thus replicated by changing the factor to 90%, 75%, 50%, and 25% of the items for two different test lengths of 30 and 75 items.

Next, varying the proportion of examinees that guessed after a certain percentage of items had been administered resulted in another set of simulations. For example, out of 60% of examinees that were flagged to start random guessing in a CAT administration, 80% started guessing after 90% of the items had been administered while 20% of the group started guessing after 75% of the items had been administered. Those percentages were then manipulated to represent other patterns of guessing behavior at each ability level.

The results from the previous steps of the study were then used to examine and describe the response patterns and their effects on the CAT administration in the following steps of the study.

In the next step of the study, a similar design was replicated. However, this time the item pool was calibrated using parameters from the November 1996 to 1998 administrations of the American Institute of Certified Public Accountants licensure examination. CATs were simulated for fixed length performance testing with and without content constraints.

Next, an adaptive mastery test was simulated where the points after which examinees' guessed were the same as performance testing. Since random guessing is expected to have a significant impact on mastery decisions for people with abilities close to the cut scores, this analysis was particularly useful. The cut-scores approximately similar to AICPA cut-scores were used for the study and the classification or mastery decision was defined as master/pass or non-master/fail.

The final step of the study involved simulating responses for variable length adaptive classification tests where the test for a given examinee depended on a certain stopping criterion. In this case, testing stopped when a required confidence level had been attained in a pass/fail decision.

#### **4.4 Item Pool Characteristics using Simulated Parameters**

The ability parameters of the examinees that were meant to take the adaptive tests were drawn from a normal distribution with mean of 0.0 and standard deviation of 1.0. The following table depicts the specifications that were used to generate item parameters:

**Table 4.1: Item Parameter Distribution**

	Distribution	Minimum	Maximum	Mean	Std. Dev.
A	Log-Normal	0.50	1.60	0.80	0.20
B	Normal	-2.50	2.50	0.00	1.20
C	Log-Normal	0.00	0.50	0.15	0.10

#### 4.5 Item Pool Characteristics using AICPA Parameters (without constraints)

In order to look at the issue of aberrance in CAT using AICPA item parameters, a careful selection of items is necessary to create a representative item pool. For similar reasons, the data from November 1996 to 1998 AICPA administrations were used. November results were used because of the similarity in the ability distributions that took the test at a particular time of the year. The November administrations were selected instead of May administrations of the tests as for November administrations results from three administrations were available to us. In other words May data were available only for 1997 and 1998 administrations of the test. Two tests with similar number of items and rather unique content were chosen for the purpose of our analyses. The two tests were Audit and Accounting and Reporting (ARE). Although analyses were performed on both tests, results from Audit will be explained thoroughly in this study while the results from ARE will be included in the appendix for readers' interest. To look at the distributions, following steps were carried out:

- a. Multiple Choice data for each administration of the two tests were cleaned for missing cases (table). The multiple-choice section for each test was composed of 75 items.
- b. Computer program BILOG was then used to calibrate the tests. Normal priors were set on the threshold parameter for better estimation.

- c. The ability estimates from phase 3 of each of the six response matrices (3 administrations x 2 tests) were read into SPSS to look at the ability distributions.
- d. Histograms of the six ability distributions were plotted. For each administration, the ability distribution was approximately normal with a mean of 0 and standard deviation 1. These distributions are depicted in figure A.1.

**Table 4.2: Ability Parameter Statistics**

Descriptive Statistics	Audit			ARE		
	1996	1997	1998	1996	1997	1998
N	50317	52292	48699	50448	52799	50554
Mean	0.0	0.0	0.0	0.0	0.0	0.0
Stdev	1.0	1.0	1.0	1.0	1.0	1.0

As shown in figure A.1 in appendix A, a very small percentage of examinees had ability levels in the tails of the distribution; majority of examinees were concentrated in the range of  $-3$  and  $+3$ . Hence the hypothesis of the similarity of the examinee ability distribution for the November administrations was supported by the analyses. Next, a few steps were performed to create a representative item pool using AICPA item parameters. Items with difficulty parameters greater than 3.0 and less than  $-3.0$  as well as items with item discrimination greater than 2.0 were deleted. Such items do not contribute in providing information about the examinees and hence were not included in the pool.

The ability distributions for all November administrations had a mean of 0 and standard deviation of 1; equating of item parameters was not deemed necessary for our

particular analyses. Item parameters from the several administrations were, therefore, combined to create a representative item pool. For audit, after combining the three administrations 223 items were available to us in the pool, while 206 items were available for the ARE. In order to create an item pool that could be sufficient for a CAT with 75 items, 600 items was considered as a “sufficient” pool size. Hence, item parameters for the available items were examined to clone the remaining items in the pool. Histograms as well as P-P probability plots of the existing item parameter distributions were carefully analyzed to simulate the remaining items. The P-P chart plots a variable’s cumulative proportions against the cumulative proportions of a number of test distributions. Such probability plots are used to determine whether the distribution of a variable matches a given distribution. If the distribution of a selected variable matches the test distribution, the points cluster around a straight line. The probability plots were analyzed for various matching distributions to decide on the most closely matched distribution. The descriptive statistics of the selected distributions are shown in table 4.2 and the actual P-P plots for those distributions are presented in figure A.2 in appendix A. Computer program CBTS was then used to generate the remaining items. A representative item pool of 600 AICPA items parameters was now available to us.

**Table 4.3: Item Parameter Statistics for Audit and ARE**

	AUDIT				ARE			
	Minimum	Maximum	Mean	Std. Dev.	Minimum	Maximum	Mean	Std. Dev.
A	0.24	1.65	0.78	0.29	0.19	1.72	0.78	0.33
B	-2.87	2.79	0.01	0.94	-2.89	2.96	0.34	1.12
C	0.07	0.46	0.24	0.07	0.08	0.50	0.25	0.09

#### 4.6 Item Pool Characteristics using AICPA Parameters with Content Constraints

The AICPA examinee booklet provides a detailed outline of the major content areas divided into several sub-areas. Audit, for example, consists of four major content areas that are in turn divided into sub-areas that range from 3 to 13 in numbers; some of these sub-areas are then refined into finer content strands. The ARE test is composed of six major content areas and each content area is then subdivided into finer strands. The examinee booklet also lists the percentage of items that are drawn from each content area while constructing the test.

For our analyses, we used the major content categories and the resulting CAT contained similar proportions of items as represented in P&P version of AICPA. These content areas and the respective percentages of items represented in the test are shown in the following table:

**Table 4.4: Content Specifications for Audit**

Content	Topic	%
1	Plan the engagement, evaluate the prospective client and engagement, decide whether to accept or continue the client/engagement and enter an agreement	40
2	Obtain and document information to form a basis for conclusions	35
3	Review the engagement to provide reasonable assurance that objectives are achieved and evaluate information obtained to reach and to document engagement conclusions	5
4	Prepare communications to satisfy engagement objectives	20

**Table 4.5: Content Specifications for Accounting and Reporting**

Content	Topic	%
1	Federal taxation --- individuals	20
2	Federal taxation --- corporations	20
3	Federal taxation --- partnerships	10
4	Federal taxation --- estates and trusts, exempt organizations, and preparers' responsibilities	10
5	Accounting for governmental and not-for-profit organizations	30
6	Managerial accounting	10

Both item pool and the test were therefore constructed to represent the respective proportions of various content categories. The number of items used for the formation of pool and the test are given below (for both tests, pool size=600; test length=75, 30):

**Table 4.6: Pool and Test Content Composition**

Content	Number of items (Audit)			Number of items (ARE)		
	Pool	Test (75)	Test (30)	Pool	Test (75)	Test (30)
1	240	30	12	120	15	6
2	210	26-27	10-11	120	15	6
3	30	3-4	1-2	60	7-8	3
4	120	15	6	60	7-8	3
5				180	22-23	9
6				60	7-8	3

The simulations were then performed for CAT with content constraints and compared to the scenarios where such constraints were not included.

#### 4.7 Mastery Testing using AICPA parameters

As mentioned earlier, the mastery tests are used to make a decision whether an



examinee passes or masters a test or fails the test. Such decisions will be called classification decisions and the passing point will be referred to as the cut-score or cut-point in the following sections. Although there are several methods available for making classification decisions, for the present study, the only method used was sequential Baye's procedure (Owen, 1975). Since the method requires the estimated ability to be compared with the latent passing score to make a classification decision, the results are influenced by the item selection only and not by the method of scoring the test (Kalohn & Spray, 1998).

According to this method, probability of mastering a test (PM) was calculated and compared with the pass decision level. If this probability was greater than the decision level, 0.5 in this case, the examinee was classified as a master. In case of variable length test, the test terminated when either of the following criteria was satisfied:

- (a)  $PM < \text{lower limit of confidence region}$
- (b)  $PM > \text{upper limit of the confidence region}$

In this case the lower limit was employed as 0.1 so the test stopped when PM was greater than 0.9 or less than 0.1. If the examinee reached the maximum limit for the number of items but none of those criteria was satisfied, the classification method was similar to the fixed length test. In other words, the decision was based on the most recent update of the mastery probability (Kalohn & Spray, 1998).

The cut-score information for AICPA tests is provided for the overall tests. The raw scores are converted to a score scale of 1-100 and the passing score for each test is 75. The classification information is also available for each of the tests. In order to obtain an estimate of the cut-score for the multiple-choice items, the classification

information was used for each administration of the test. The derived cut-score information from each ability distribution is presented below:

**Table 4.7: Derived Cut-Scores**

	Audit			ARE		
	1996	1997	1998	1996	1997	1998
<b>Derived cut-score</b>	0.52	0.54	0.49	0.52	0.64	0.57

The average derived cut-scores were, therefore, taken as 0.52 and 0.58 for Audit and ARE respectively on the IRT ability scale.

The mastery tests, like performance (achievement) tests were also being simulated at two different test lengths. First, fixed length and then variable length tests were administered to the 1200 examinees. In case of variable length CAT, the examinees could take a minimum of 25 items and a maximum of 45 items. Those limits were chosen to depict half of the test length for paper and pencil version of the tests.

A significant aspect of these analyses was to look at the effect of rushed guessing on the classification decisions in a licensure examination. The effect of guessing on the ability estimation as well as the accuracy of classification decisions was analyzed for both fixed and variable lengths tests.

#### 4.8 Analyses

In order to look at the differences between true and estimated ability estimates, Root Mean Square Error (RMSE) and Bias indices were computed. Here, root mean

square index is defined as the standard deviation of estimated ability around true ability.

Bias, on the other hand provides us with a sense of direction of the estimated abilities relative to the truth. Following is the mathematical representation of the two indices:

$$RMSE = \sqrt{\frac{\sum_{j=1}^N (\hat{\theta}_j - \theta)^2}{N}}$$

$$BIAS = \frac{\sum_{j=1}^N (\hat{\theta}_j - \theta)}{N}$$

Here, N is the total number of examinees. The values for those indices are presented in both tabular and graphical formats for various levels of aberrance. For each administration of the test, average test information was computed at various ability levels for different response patterns. This provided us with an idea of how well the test was targeted to the examinee ability levels. For each aberrant condition, various plots were produced for the estimated ability against the items currently administered. This was repeated for several examinees at a variety of ability levels. Examinees with the same true abilities (which is frequent) were considered as the replications of the estimation at a particular ability level thus giving us a clear picture of the estimation process.

It would also be helpful to look at an index of pool utilization. Although a lenient exposure control method was used, it'd be helpful to look at the effect of aberrance on exposure rates and thus the pool utilization. We could hypothesize that greater aberrance could lead to increased Skewness in the exposure rates. One such index was proposed by

Chang & Ying (1999) is a chi-square index of pool utilization that provides a measure of Skewness of exposure rates (Robin, 2000). The index is defined as, (to be included)

$$\chi^2 = \frac{\sum_{j=1}^N (er - L/N)}{N}$$

where  $er$  is the exposure rate,  $L$  is the number of items administered,  $N$  is the number of items in the pool. Although, the index cannot be relied upon in isolation, that is, without looking at exposure rates, it's used in this study to look at the general patterns of pool utilization for different guessing behaviors.

The information provided by the pool that was available for an examinee before the item selection began, was also examined.

## CHAPTER 5

### RESULTS

#### 5.1 Results Based on Simulations

The first round of simulations was carried out to generate responses on computer adaptive tests for 1200 examinees at two test lengths of 30 and 75 items. An item pool of 600 items was simulated for these analyses. The examinees were made to guess after a certain percentage of items had been administered to look at the effect of guessing on the response patterns and the final ability estimates. The examinees were made to guess after 90% of items had been administered to simulate examinees that guess later in the exam. On the other hand, guessing after 25% of test administration depicts the response patterns for examinees that are extremely slow test takers or have very low ability and start guessing very early in the test.

Figure B.1.a demonstrates a computer adaptive test administration for a low ability examinee when he/she started guessing at several points in time, ranging from very early to later in the test. Figures B.1.b and B.1.c show the same results for examinees with middle and higher levels of proficiency. For the purpose of these analyses, theta levels of  $-1.83$  and  $1.83$  were arbitrarily chosen to depict lower and higher levels of ability, respectively. Theta level of  $0.1$  was chosen to simulate responses for examinee with middle ability level. The following table displays the ability levels and the midpoint of the corresponding ability interval (the interval size was smaller around the mid-ability compared to the tails of the population to simulate a normal distribution of examinees).

**Table 5.1: Ability Levels**

Ability Levels	Mid-Point	Ability	Mid-Point
1	-1.83	7	0.10
2	-1.15	8	0.31
3	-0.80	9	0.54
4	-0.54	10	0.80
5	-0.31	11	1.15
6	-0.10	12	1.83

Figures B.2.a, B.2.b, and B.2.c demonstrate the same analyses, however, the first set of analyses was performed for a test length of 30 while the second set for a test length of 75 items. The test length of 75 items reflects the test lengths for the various sub-tests for AICPA exam (the four sub-tests consist of 60 to 75 items). The plots indicate the estimated ability and item difficulty (vertical axes) after each item has been administered (horizontal axis). The vertical dashed line indicates the point after which an examinee starts guessing while the horizontal line is drawn across the true ability estimate of the examinee.

The results showed that the ability estimation considerably improved when the test length was increased from 30 to 75 items. An important finding in both cases was that the adverse effects of guessing were significant for middle and high ability examinees. Those effects were highly noticeable when the examinees started guessing at the earlier stages of the test. If we looked at the examinees that started guessing halfway through the test, the difference between the true and final estimated ability was a fraction of a point for a low ability examinee while one and two point difference was observed for middle and high ability examinees respectively. Increasing the test length improved the

estimation for the high ability examinee when he/she started guessing at a later point in the test. The test length didn't prove to have much of an effect on the estimation in general once the examinees started to guess. This result is demonstrated in figure B.3 where root mean squared error in the ability estimation over 1200 examinees are plotted against the guessing points for the two test lengths. Increasing the test length actually resulted in higher RMSE and Bias indices.

The effect of guessing on the actual examinee responses is shown in figure B.4. The figure shows an example of the way a pattern of responses could change when the examinee guesses. Here the true ability of the examinee was moderately high, hence a number of correct responses were obtained when the examinee did not guess under the "null" condition. The responses were adversely affected (more 0s than 1s) when the examinee rushed after a certain number of items had been administered.

One of the hypotheses that we could also formulate from our knowledge of the CAT is that if a large number of examinees are simultaneously taking a CAT from the same item pool, ability estimation for the group could be affected. Since the utilization of the pool is disturbed and distracted from the way it was supposed to be utilized, the quality of estimation could be affected in general. Various proportions of examinees that took the test about the same time were thus made to guess at various points in the test. Figure B.5 illustrates the fact when 60% of the examinees were flagged or were made to guess. For those flagged examinees, several scenarios were simulated as described below:

Scenario A: 20% of flagged examinees guess after 25% of items have been administered; 80% guess towards the end (after 90% of items have been administered)

Scenario B: 20% of flagged examinees guess after 50% of items have been administered; 80% guess towards the end

Scenario C: 20% of flagged examinees guess after 75% of items have been administered; 80% guess towards the end

Scenario D: All examinees guess towards the end of the test

For all the above scenarios, the root mean square errors in the final ability estimates were plotted for each situation as shown in figure B.5. The results reiterate the fact of how rushed guessing starting at early stages in a CAT could influence the final ability estimates. A very large population of examinees who guess towards the end of the test could also influence the pool utilization such that the very few examinees that started guessing early might end up with very poor ability estimates. This hypothesis could be true if there was a lack of high quality easy items available for guessing examinees. The hypothesis will be examined more carefully in the next phase of study.

Figure B.6.a and B.6.b depict the values for average test information that the test would provide assuming that the examinees had rushed into guessing. Out of 1200 examinees that took the CAT about the same time, all of them guessed at one point or the other in figure B.6.a. On the other hand, figure B.6.b shows that only 30% of the 1200 examinees guessed at those points. As expected the test information is significantly affected when a higher number of people guess, however the differences between the two situations are drastic when guessing took place early in the test. The small percentage of population who guessed earlier had a serious impact on the average test information. The amount and pattern of guessing across the 12 ability levels stayed almost similar in both situations when the guessing took place later in the test.



## 5.2 Results for Proficiency Testing using AICPA parameters

The results for CAT simulated with AICPA parameters were based on the analyses performed for the simulated parameters. The root mean squared errors (RMSE) were used to examine the estimation accuracy. Figure C.1 depicts the distribution of examinees falling in a certain ability interval based upon true and estimated ability. A significant drop was observed in the number of examinees in the higher ability intervals as they started to guess. Same analyses were repeated for other types of tests (mastery etc). The figures for those tests are shown together so that readers can compare the figures when a reference is being made to those tests later in this chapter.

Figure C.2.a through C.2.c show the Root Mean Squared Errors plotted against the various guessing behaviors for various ability levels at test lengths of 30 and 75 items. Figure C.2.d indicates the overall RMSE while the rest depict the plots for low, medium and high ability examinees. For a 75-item CAT, the error in estimation increased from 0.2 when there was no guessing to 0.3 when guessing was introduced towards the end, 0.7 when guessing began after 75% of the test was administered, 2.5 after 50% and 3.6 when guessing began very early. For a 30-item CAT, these values were 0.3 for no guessing situation, 0.4 when examinees guessed towards the end, 0.8, 2.0 and 3.3 for the respective guessing behaviors. Also shown in figure C.2.d, the RMSE values were slightly higher for a 30-item test than 75-item test when guessing was introduced in the later part of the test while lower when guessing began earlier. When the RMSE errors were examined for examinees at various ability levels, it was found that the errors followed similar patterns for the two test lengths. An exception to this was the case of low ability examinees for

whom; the RMSE values constantly remained higher for the 30-item test. The following table presents the values of RMSE for the three ability levels:

**Table 5.2: Error in Estimation (RMSE) for Audit Sub-test**

Guessing	RMSE for a 75-item CAT				RMSE for a 30-item test			
	Overall	Low	Medium	High	Overall	Low	Medium	High
25%	3.61	1.45	3.67	5.40	3.28	1.69	3.14	4.70
50%	2.48	1.22	2.51	3.64	2.03	1.27	2.05	2.89
75%	0.66	0.49	0.68	0.94	0.75	0.74	0.77	0.88
90%	0.29	0.35	0.26	0.41	0.39	0.50	0.36	0.44
NG	0.19	0.30	0.18	0.20	0.27	0.36	0.26	0.24

The RMSE values were very similar at each ability level when the examinees did not rush into guessing. The errors were constantly higher for the high ability examinees and their differences from the middle and low ability examinees increased as examinees guessed early on. The RMSEs for examinees with middle ability levels were lower than the high ability examinees but higher than the low ability examinees in terms of error in estimation.

Bias in estimates is represented in table 5.3 (see figure C.3). The table shows that

**Table 5.3: Bias in Estimates for Audit subtest**

Guessing	Bias for a 75-item CAT				Bias for a 30-item test			
	Overall	Low	Medium	High	Overall	Low	Medium	High
25%	1.83	1.13	1.89	2.31	1.74	1.23	1.72	2.13
50%	1.45	1.04	1.50	1.83	1.30	1.02	1.31	1.59
75%	0.75	0.59	0.75	0.95	0.75	0.70	0.74	0.88
90%	0.44	0.37	0.39	0.60	0.48	0.52	0.44	0.59
NG	0.09	0.22	0.04	0.09	0.13	0.26	0.09	0.17

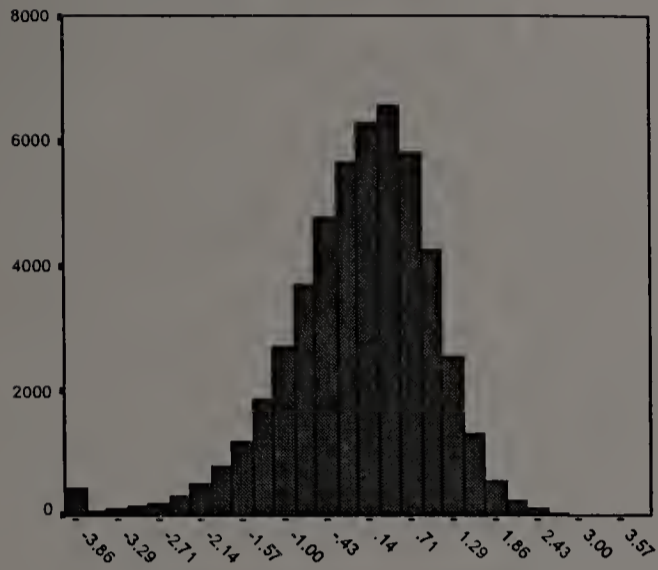
the bias increased significantly as soon as the guessing was introduced. Although, the RMSE values were negligibly small when examinees guessed towards the end, the bias in

APPENDIX A

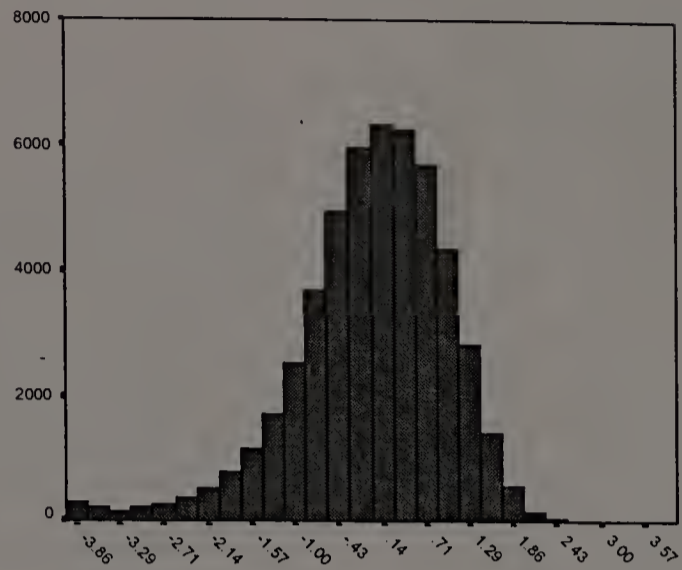
ABILITY AND ITEM PARAMETER DISTRIBUTIONS

# Ability Distributions for November Administrations of Audit and ARE

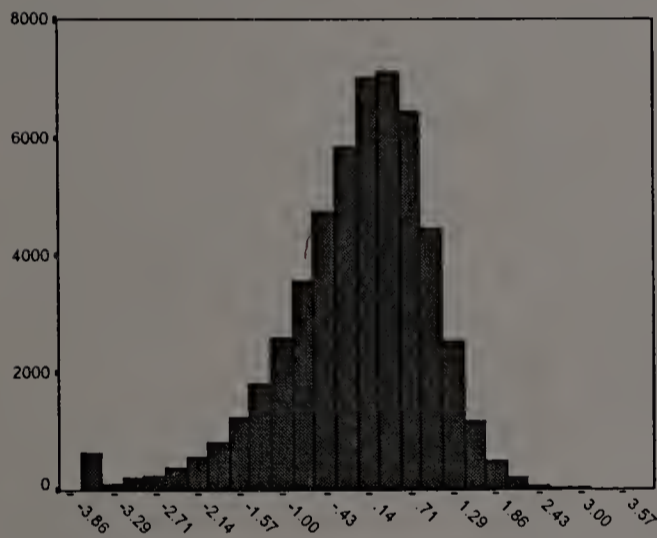
(1) 1996 Audit



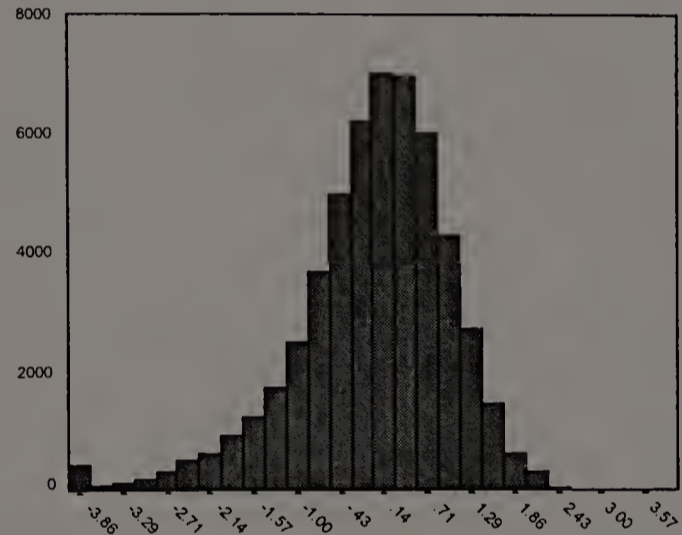
(2) 1996 ARE



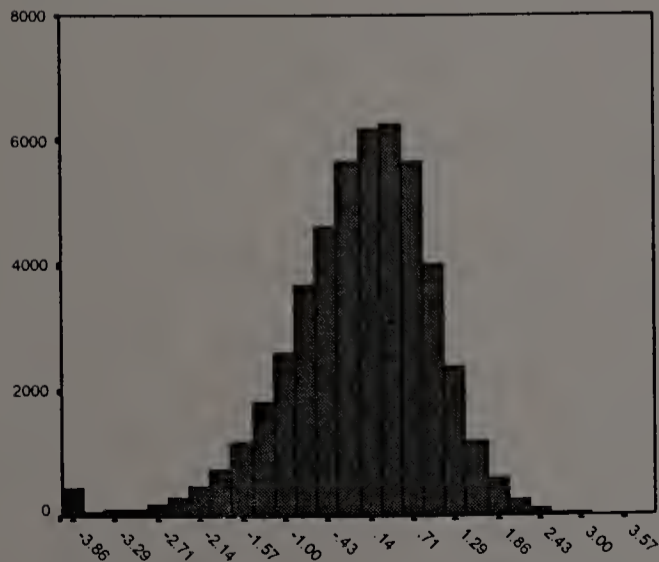
(3) 1997 Audit



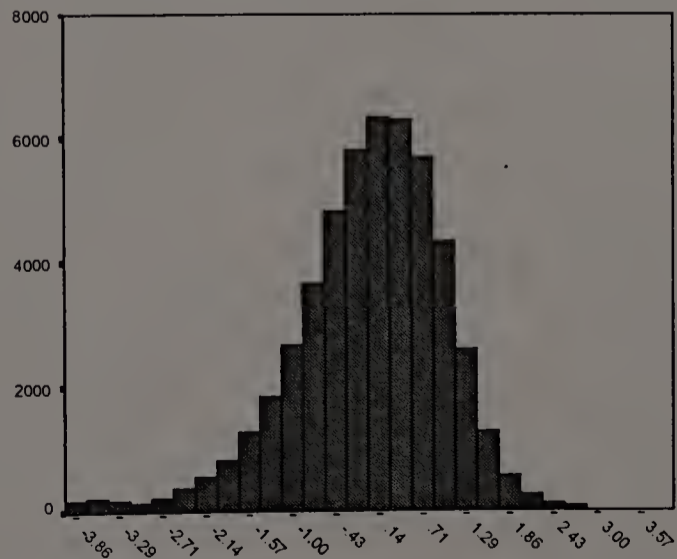
(4) 1997 ARE



(5) 1998 Audit

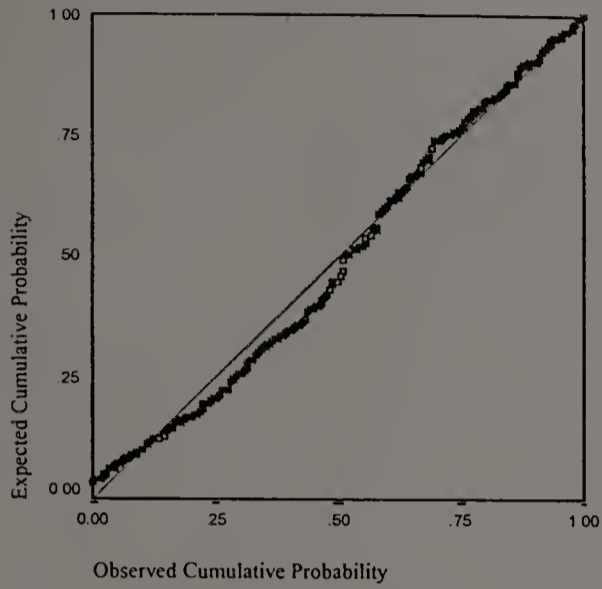


(6) 1998 ARE

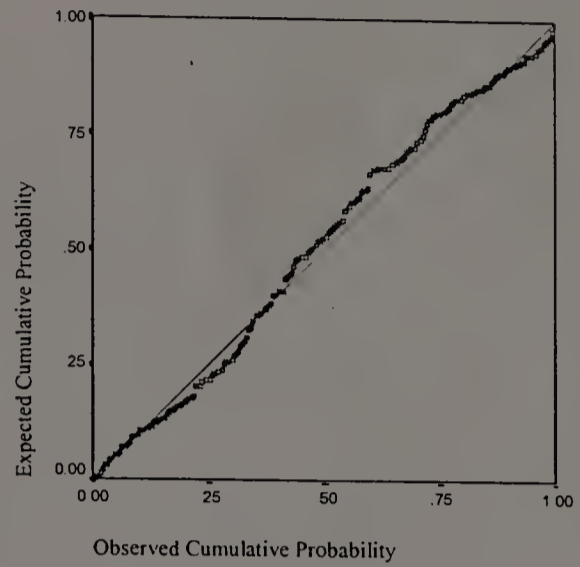


# P-P Plots for AICPA Item Parameters

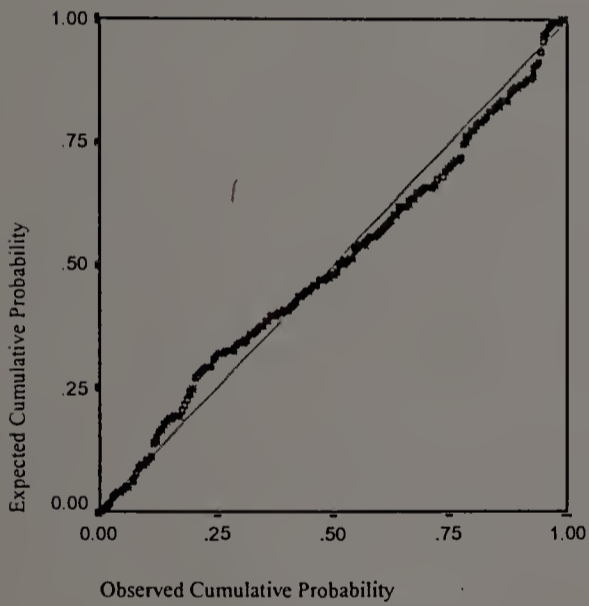
(7) Normal P-P for Audit a



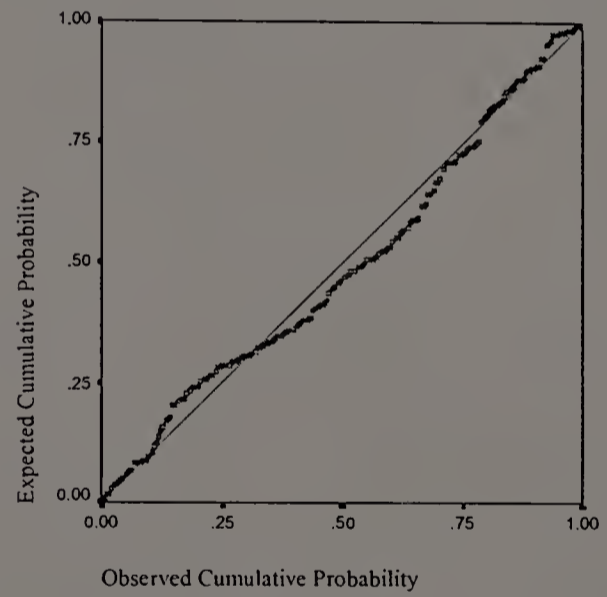
(8) Log-Normal P-P for ARE a



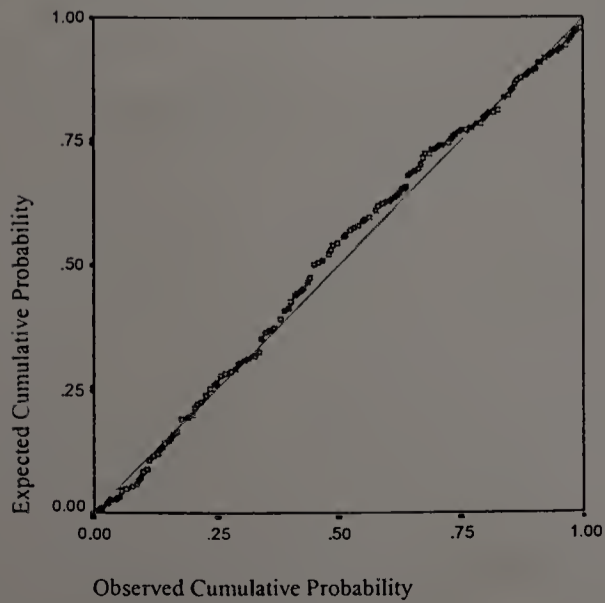
(9) Normal P-P for Audit b



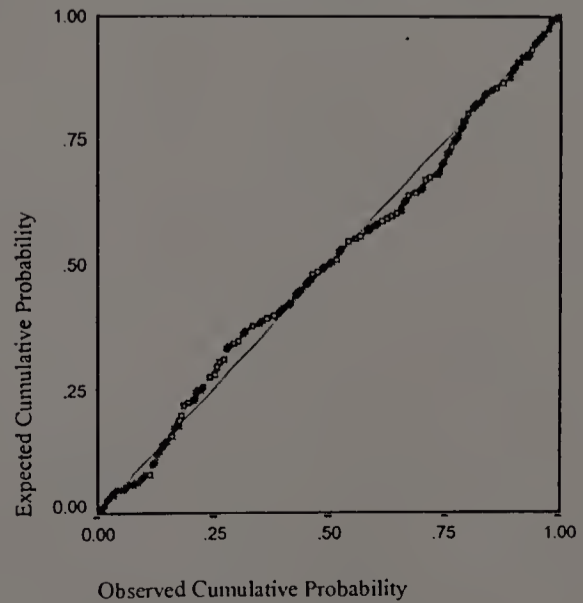
(10) Normal P-P for ARE b



(11) Normal P-P for Audit c



(12) Log-Normal P-P for ARE c

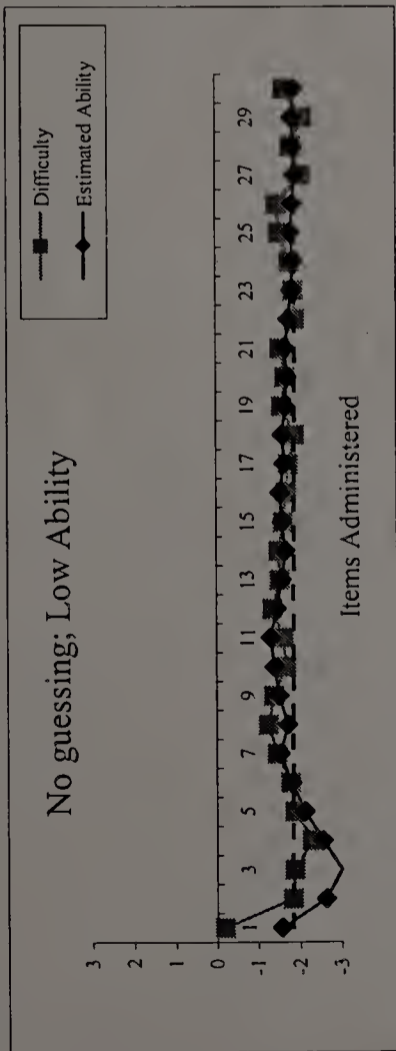


APPENDIX B

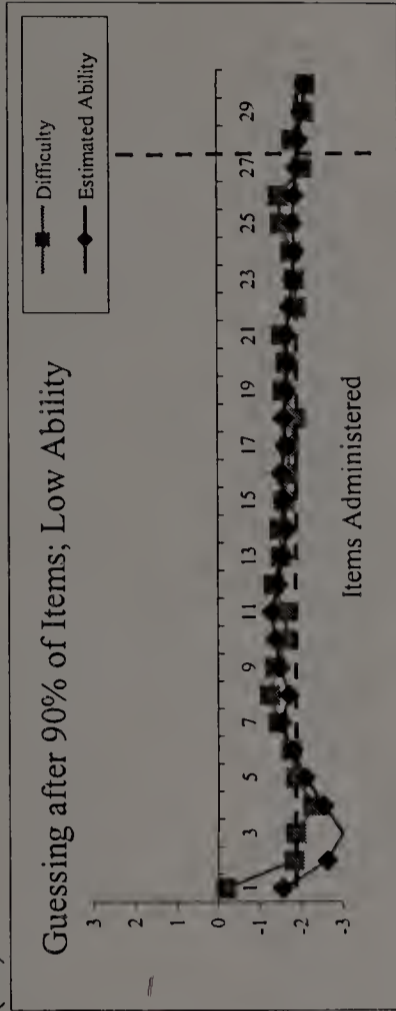
RESULTS USING SIMULATED ITEM PARAMETERS

CAT Administration for a Low Ability Examinee who Guesses at a Certain Point in the Test (30 items)

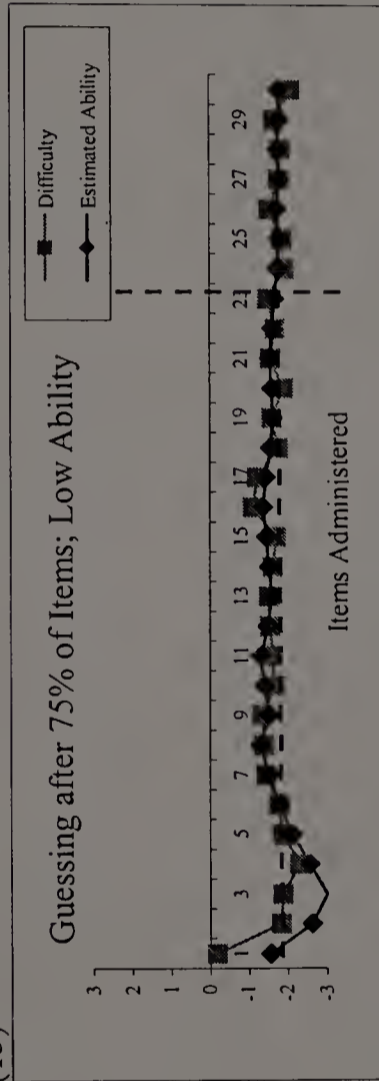
(13)



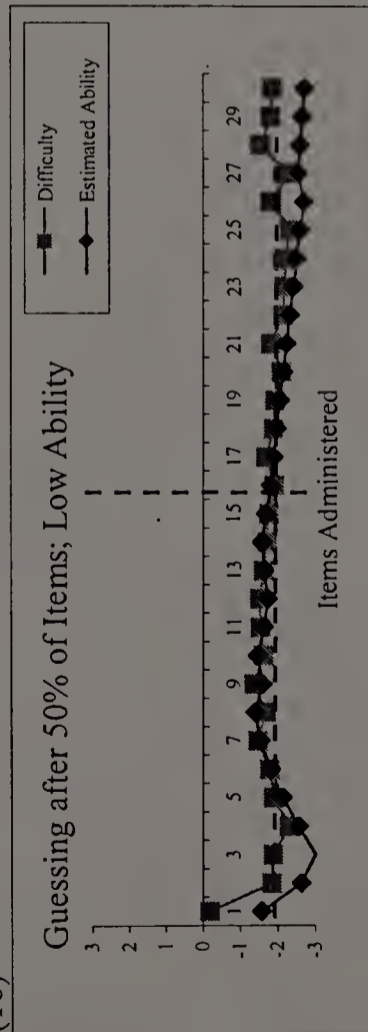
(14)



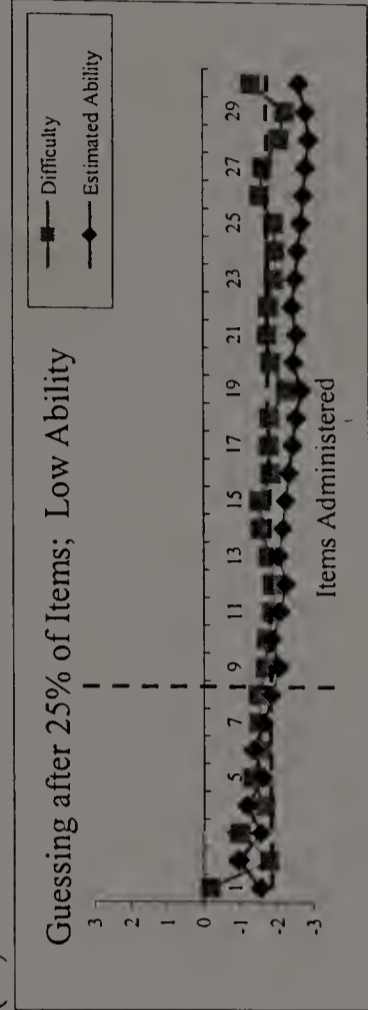
(15)



(16)

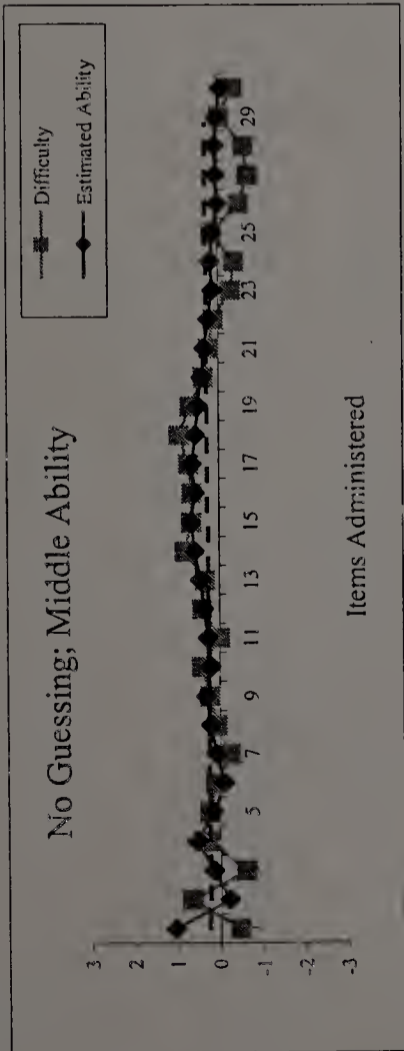


(17)

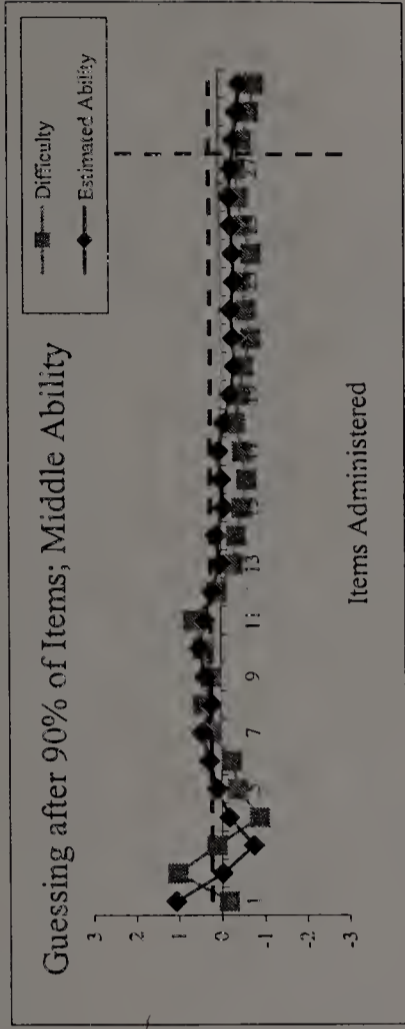


CAT Administration for a Middle Ability Examinee who Guesses at a Certain Point in the Test (30 items)

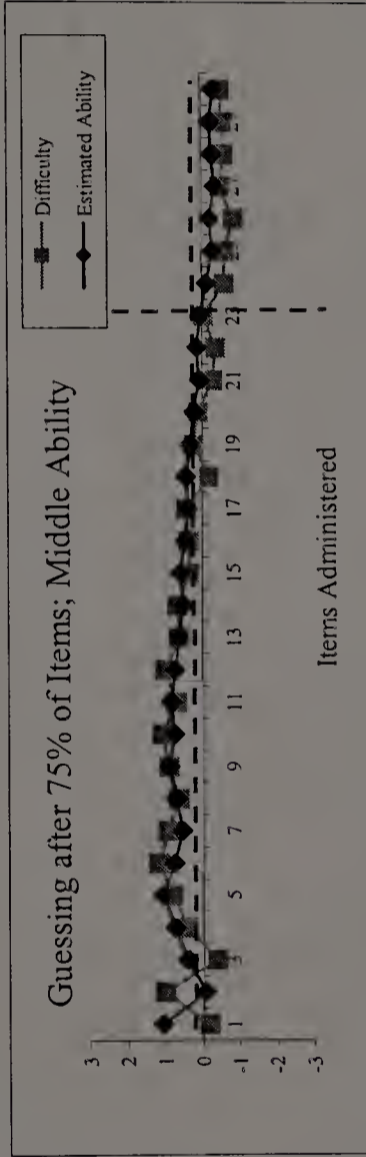
(18)



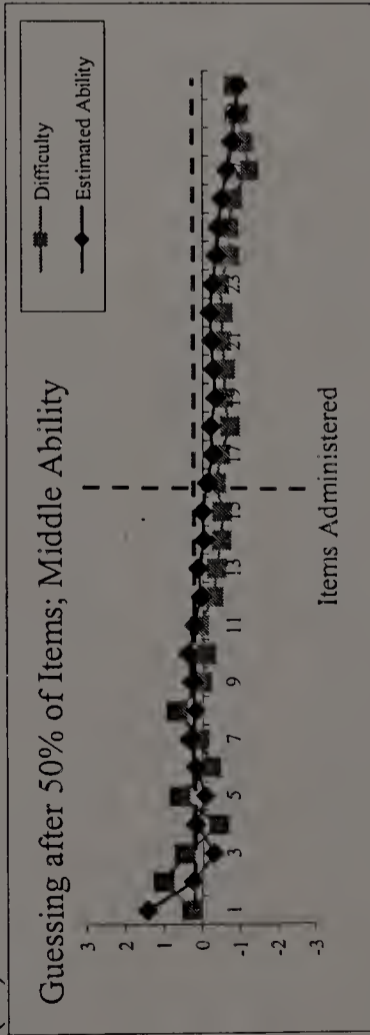
(19)



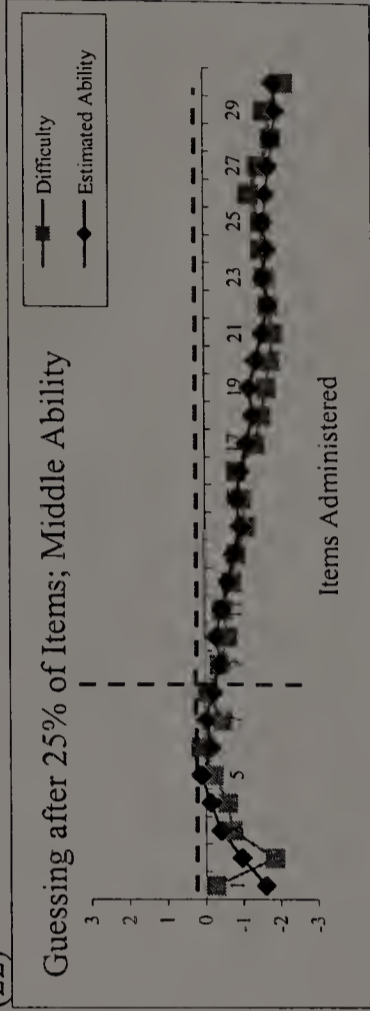
(20)



(21)

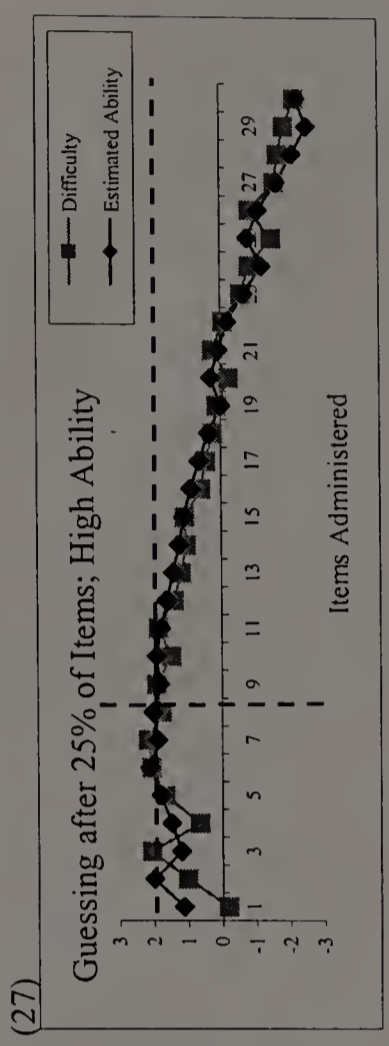
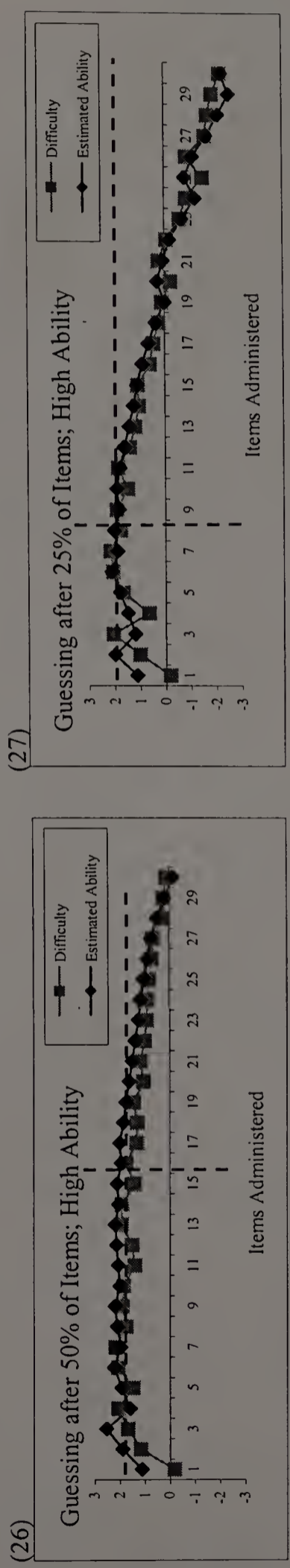
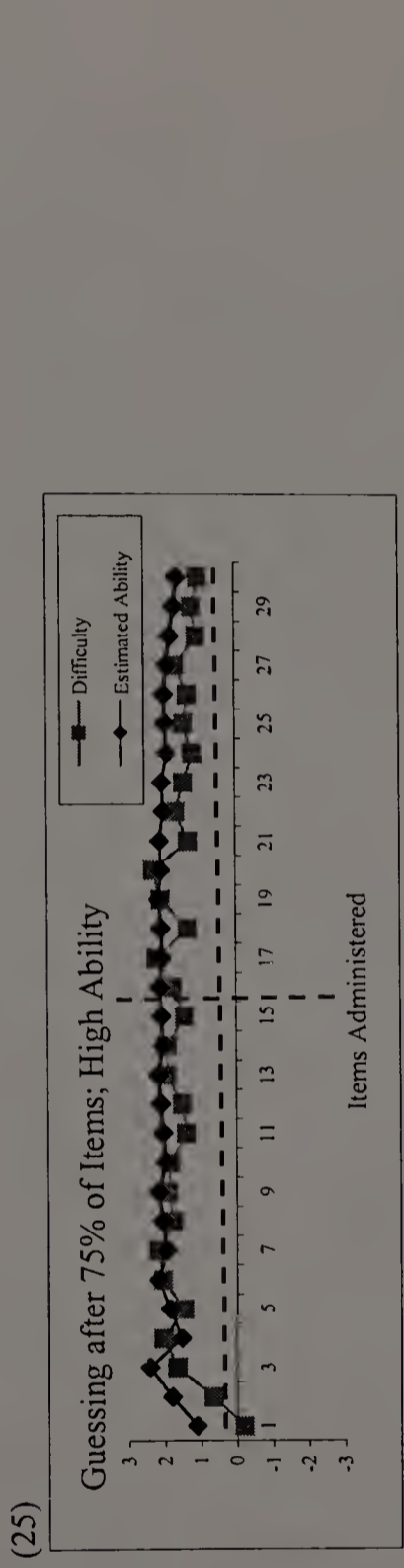
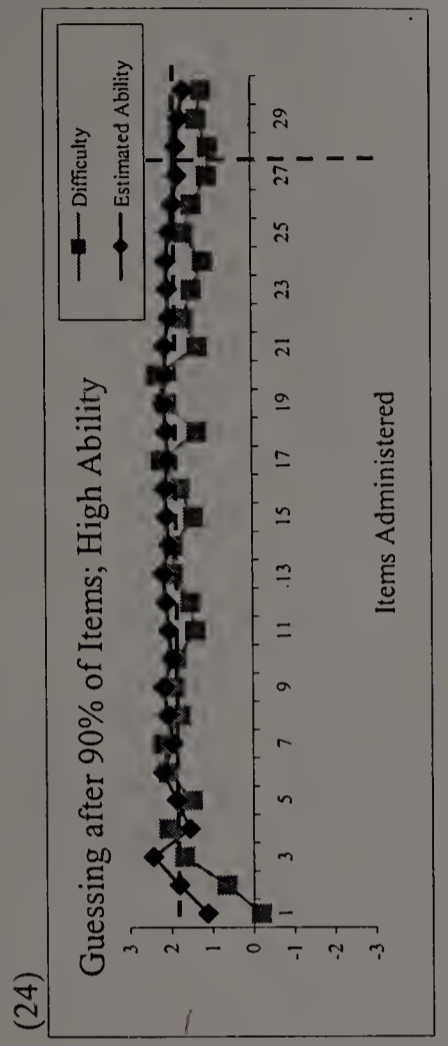
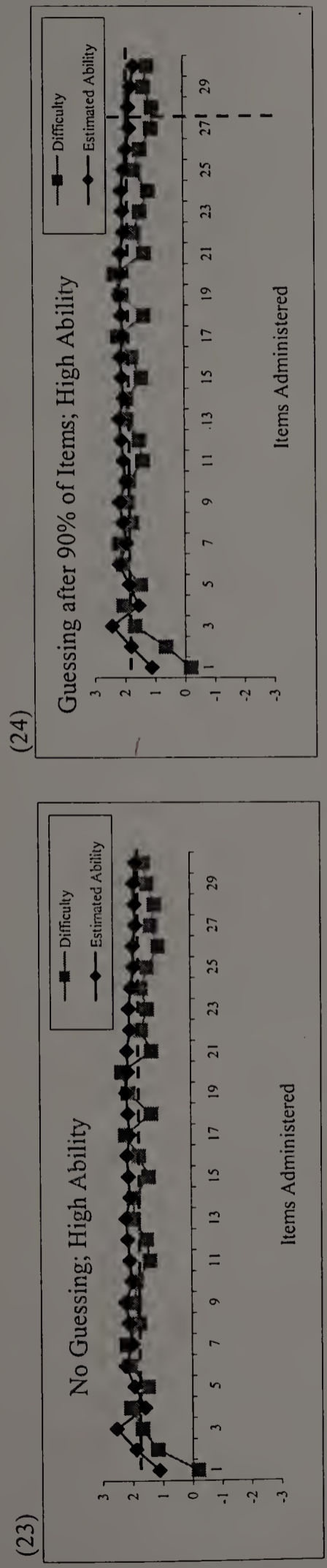


(22)



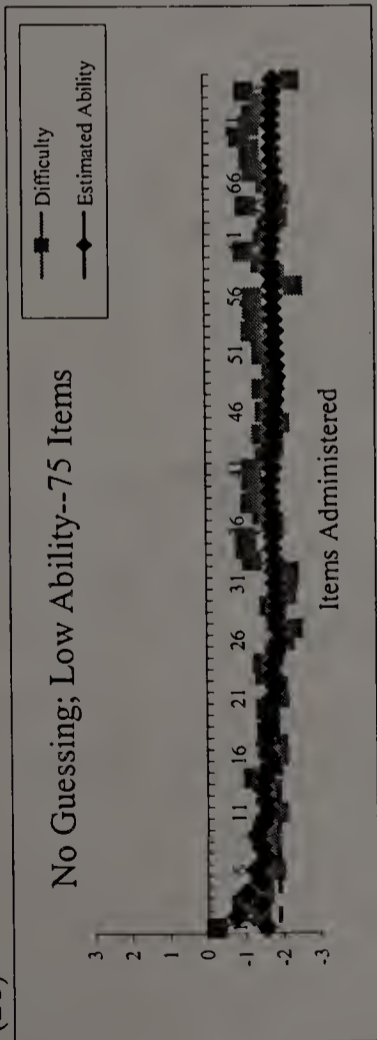


CAT Administration for a High Ability Examinee who Guesses at a Certain Point in the Test (30 items)

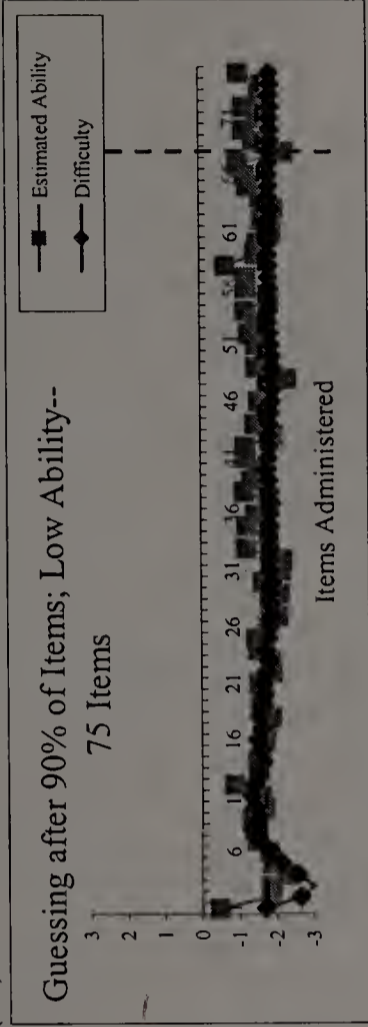


CAT Administration for a Low Ability Examinee who Guesses at a Certain Point in the Test (75 items)

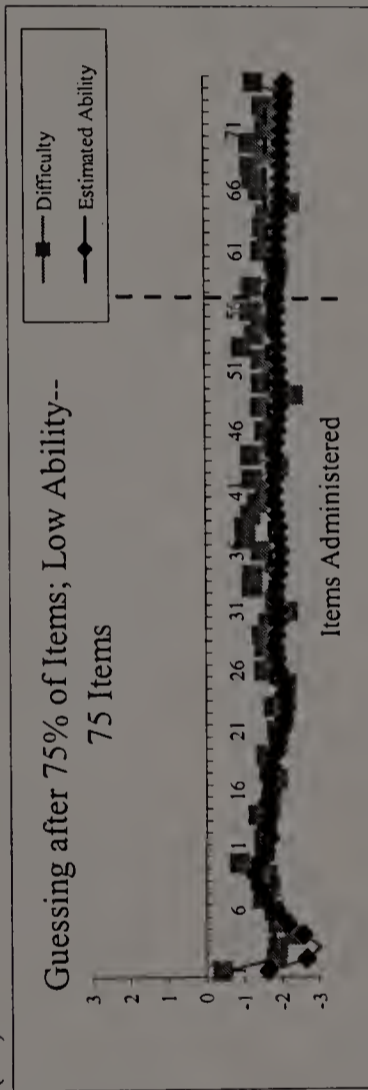
(28)



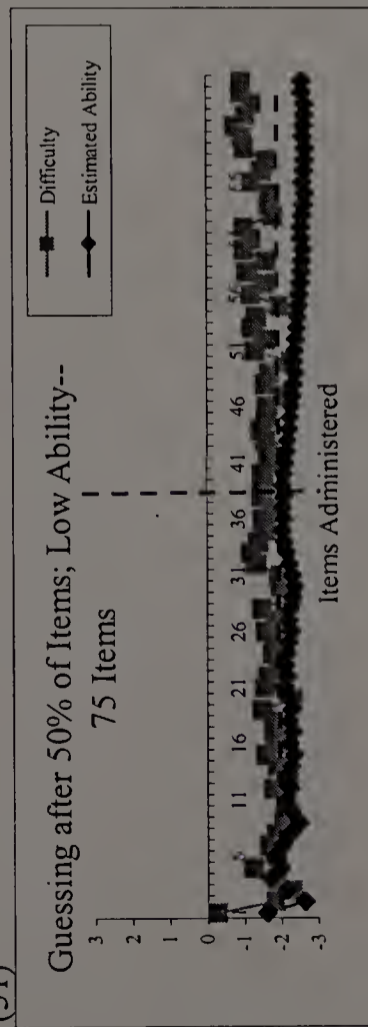
(29)



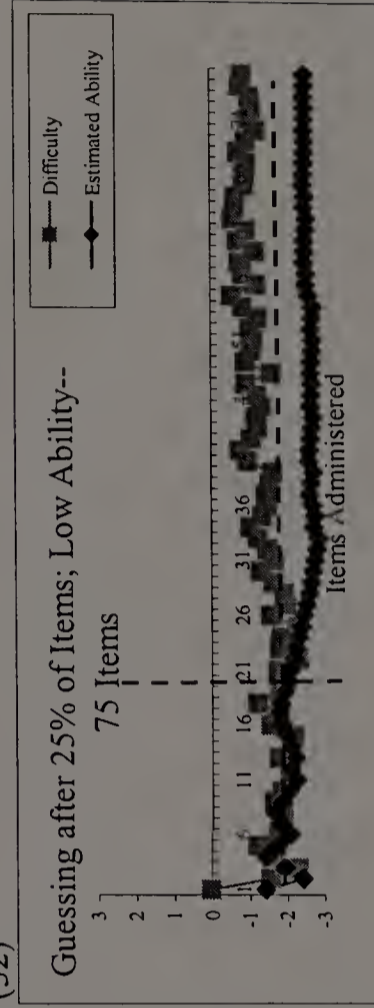
(30)



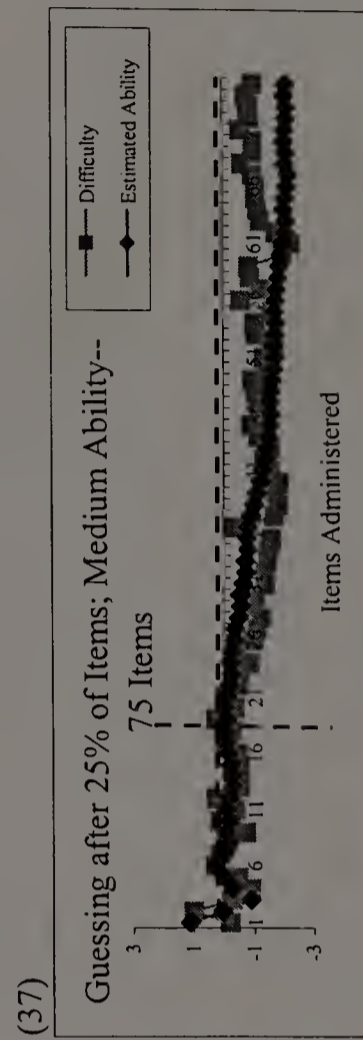
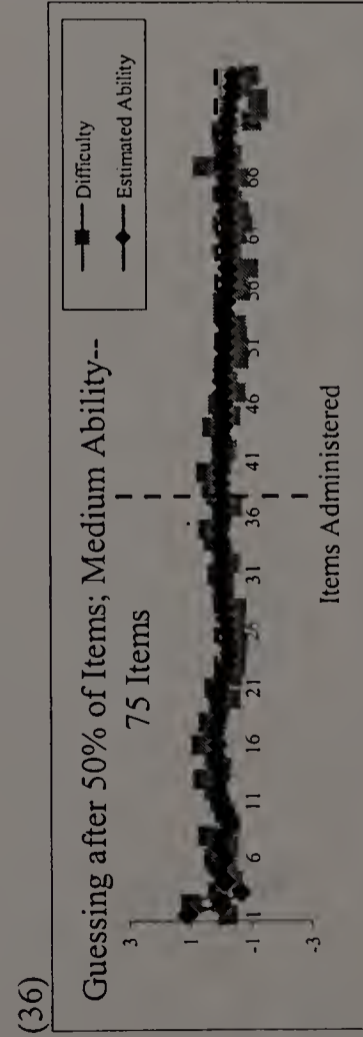
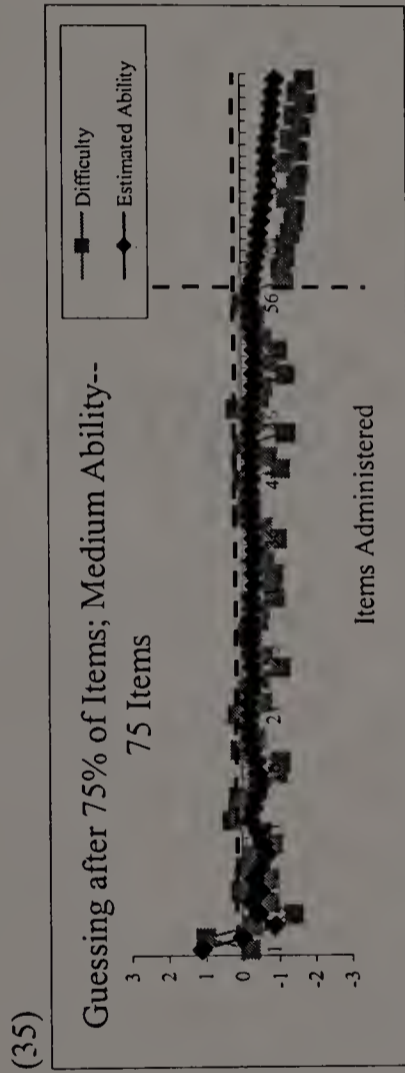
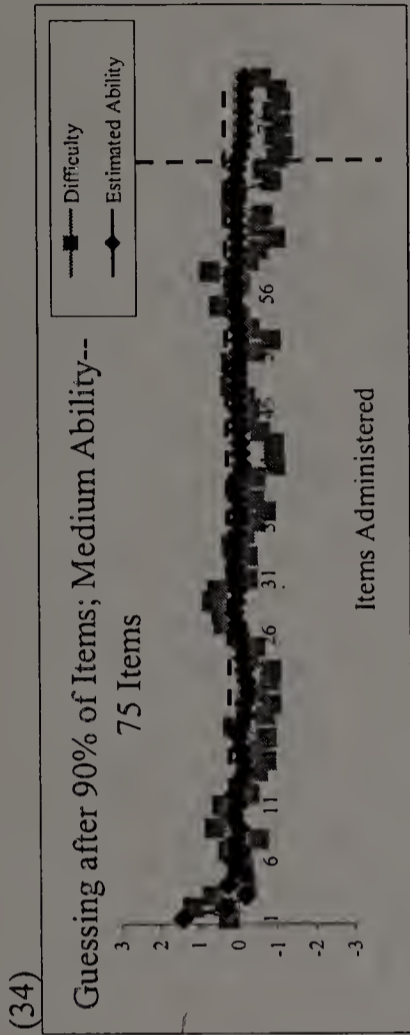
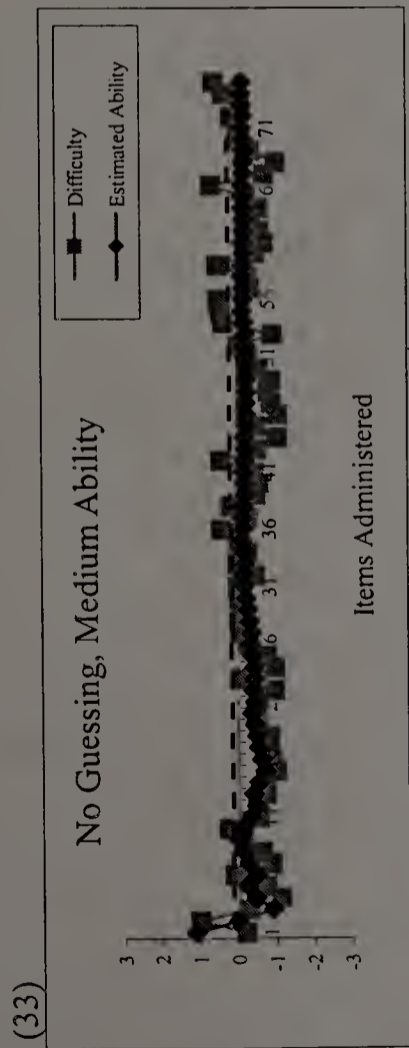
(31)



(32)

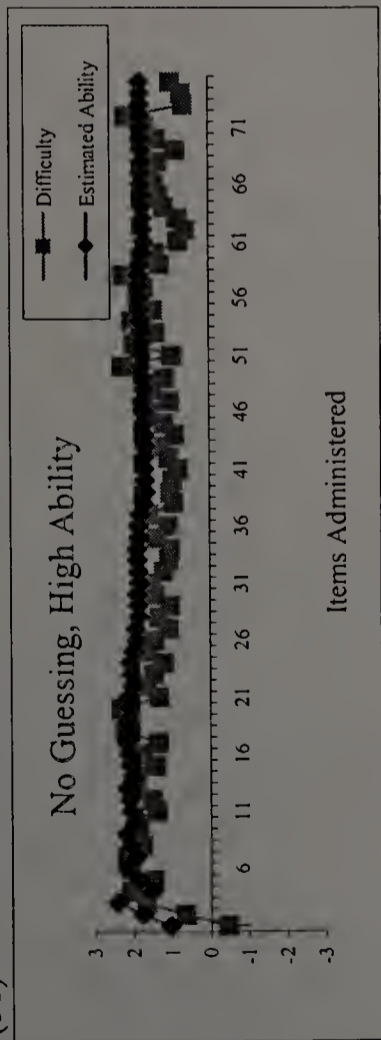


CAT Administration for a Medium Ability Examinee who Guesses at a Certain Point in the Test (75 items)

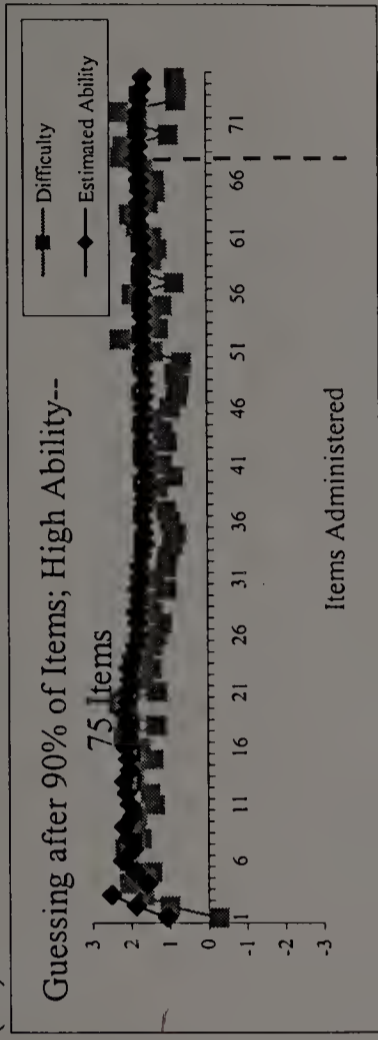


CAT Administration for a High Ability Examinee who Guesses at a Certain Point in the Test (75 items)

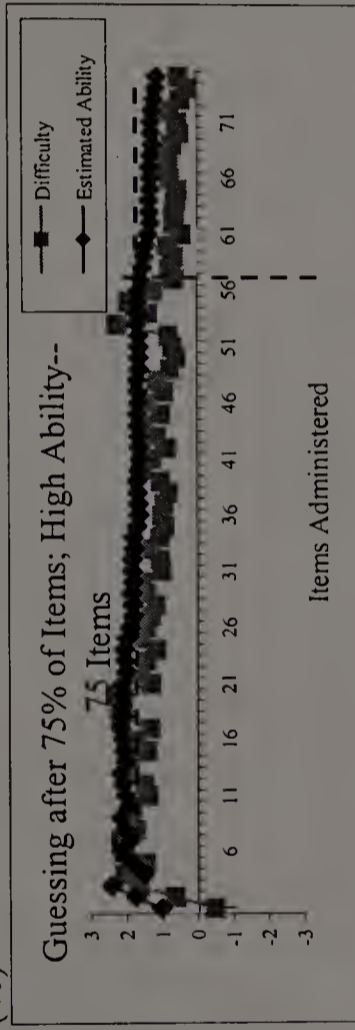
(38)



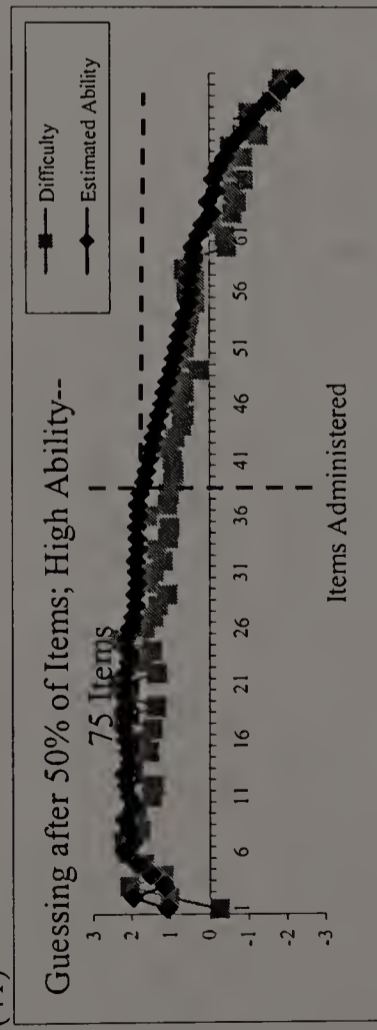
(39)



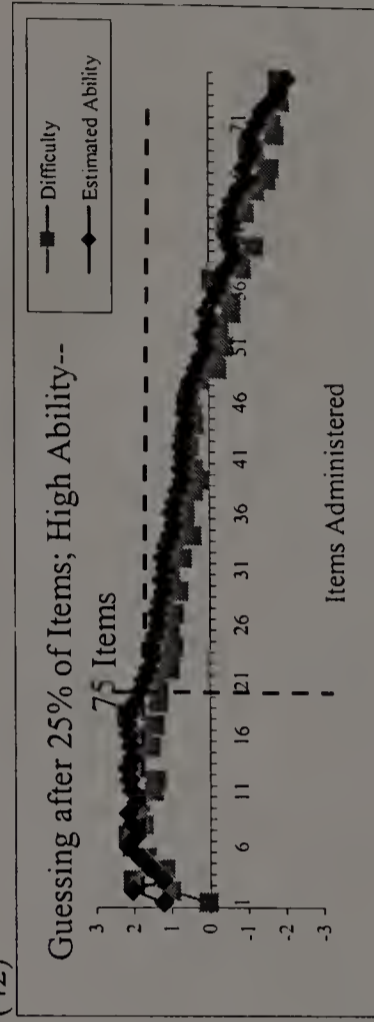
(40)



(41)

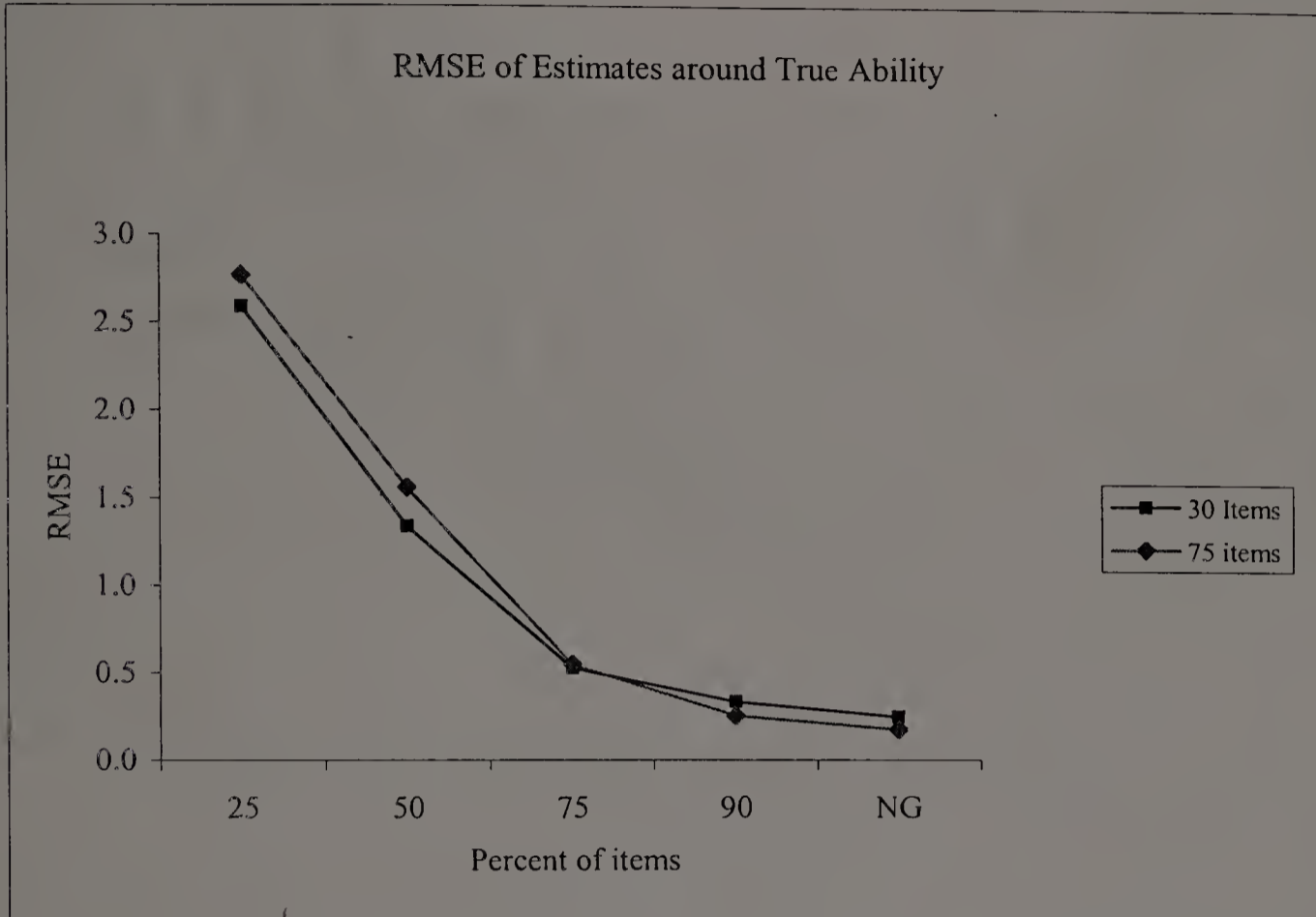


(42)



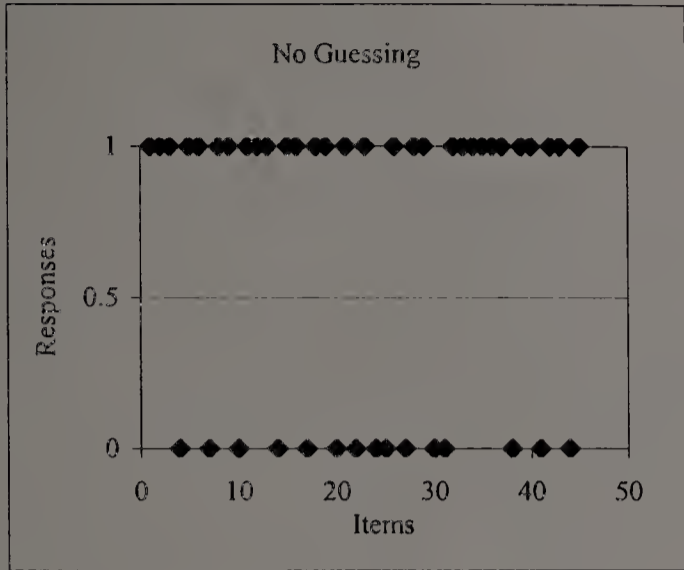
# Error in Ability Estimation for Various Guessing Behaviors at Two Test Lengths

(43)

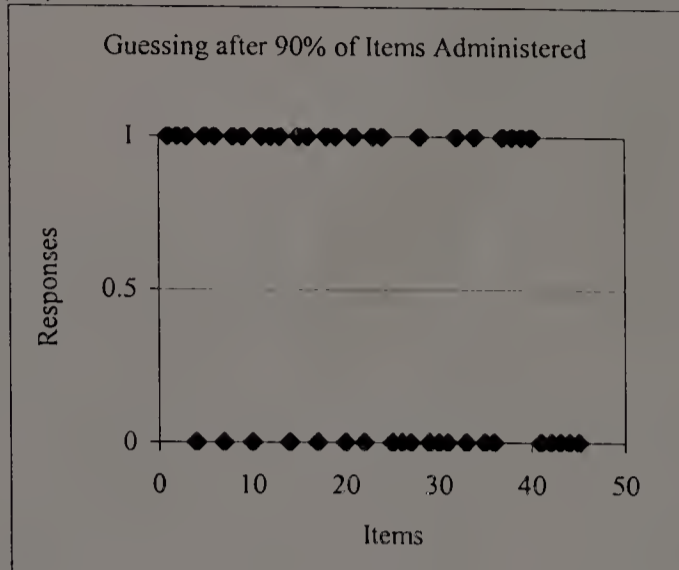


# Responses for a High Ability Examinee for Various Guessing Behaviors

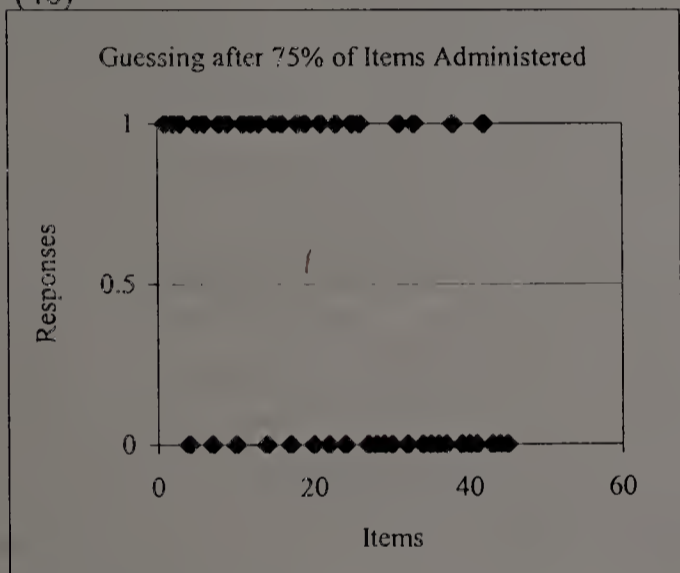
(44)



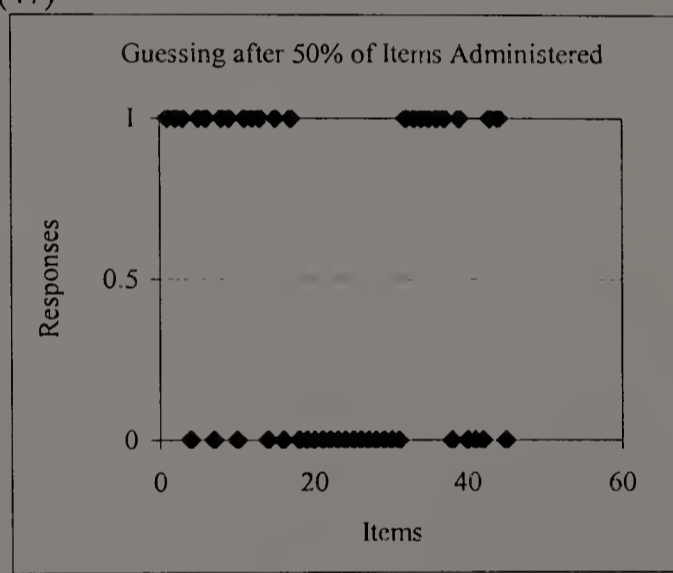
(45)



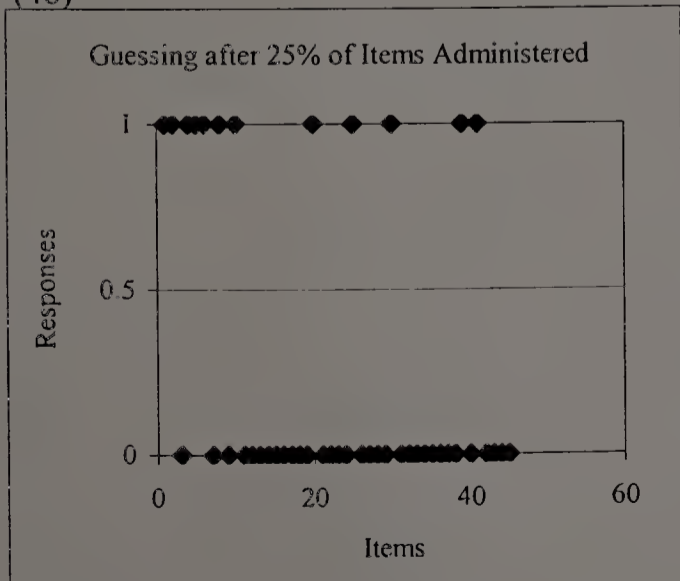
(46)



(47)

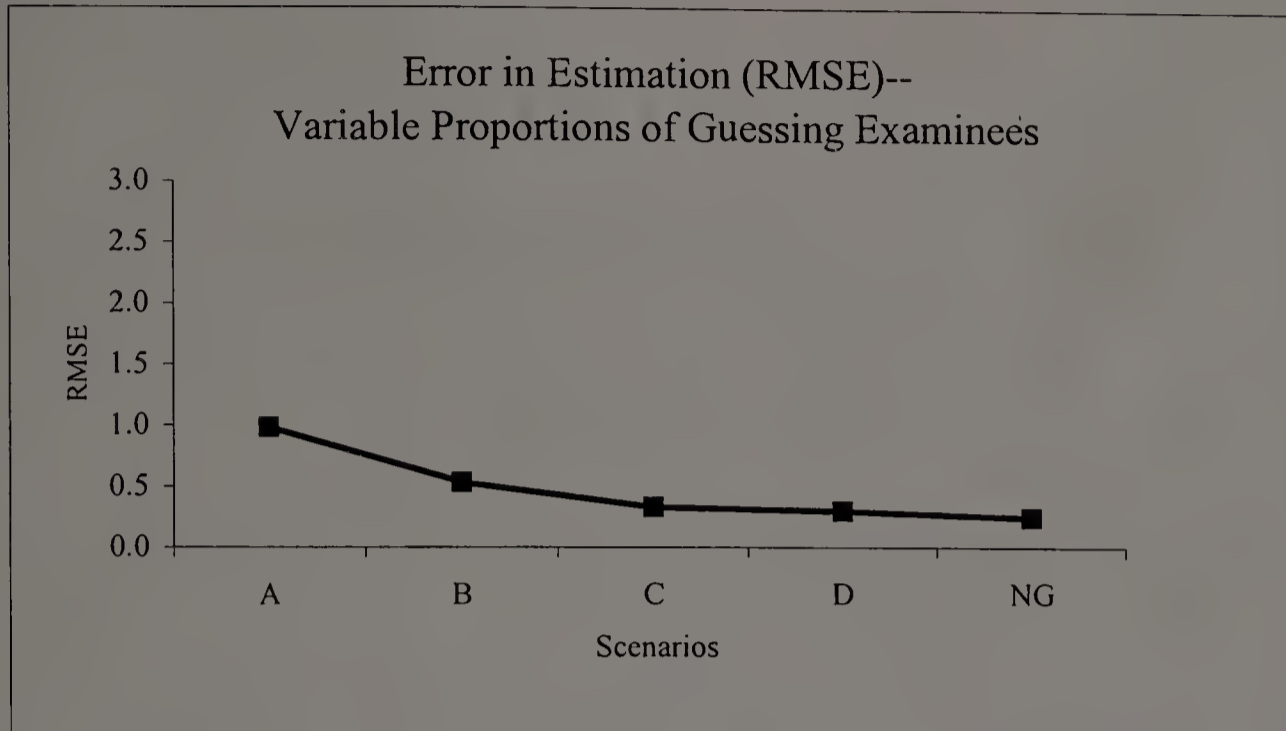


(48)



## Error in Estimation due to Variable Proportions of Guessing Examinees

(49)



Note: Guessing introduced for 60% of examinees (flagged) at each ability level

Scenario A: 20% of flagged examinees begin guessing after 25% of items had been administered  
80% guess towards the end (after 90% of items have been administered)

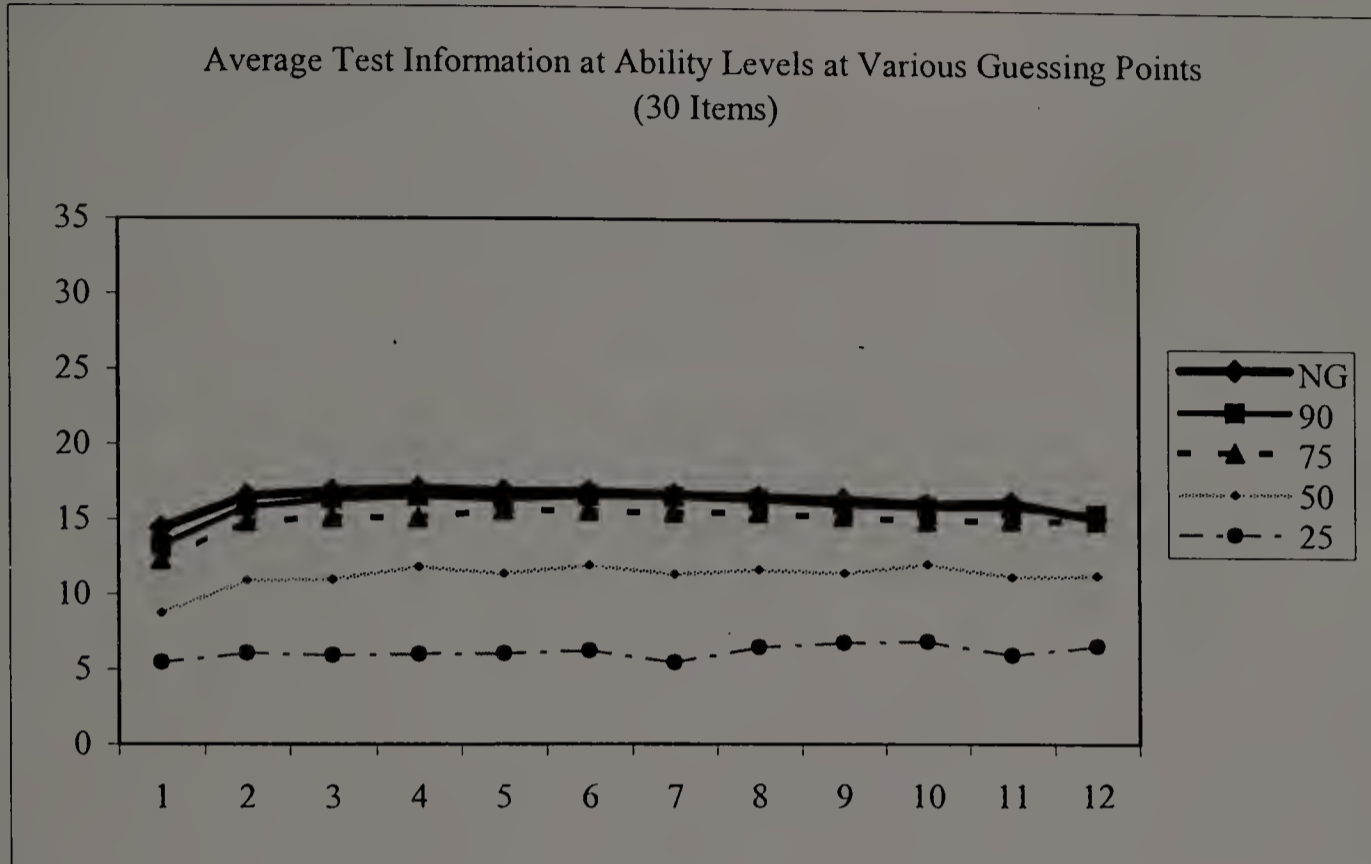
Scenario B: 20% of flagged examinees begin guessing after 50% of items had been administered  
80% guess towards the end

Scenario C: 20% of flagged examinees begin guessing after 75% of items had been administered  
80% guess towards the end

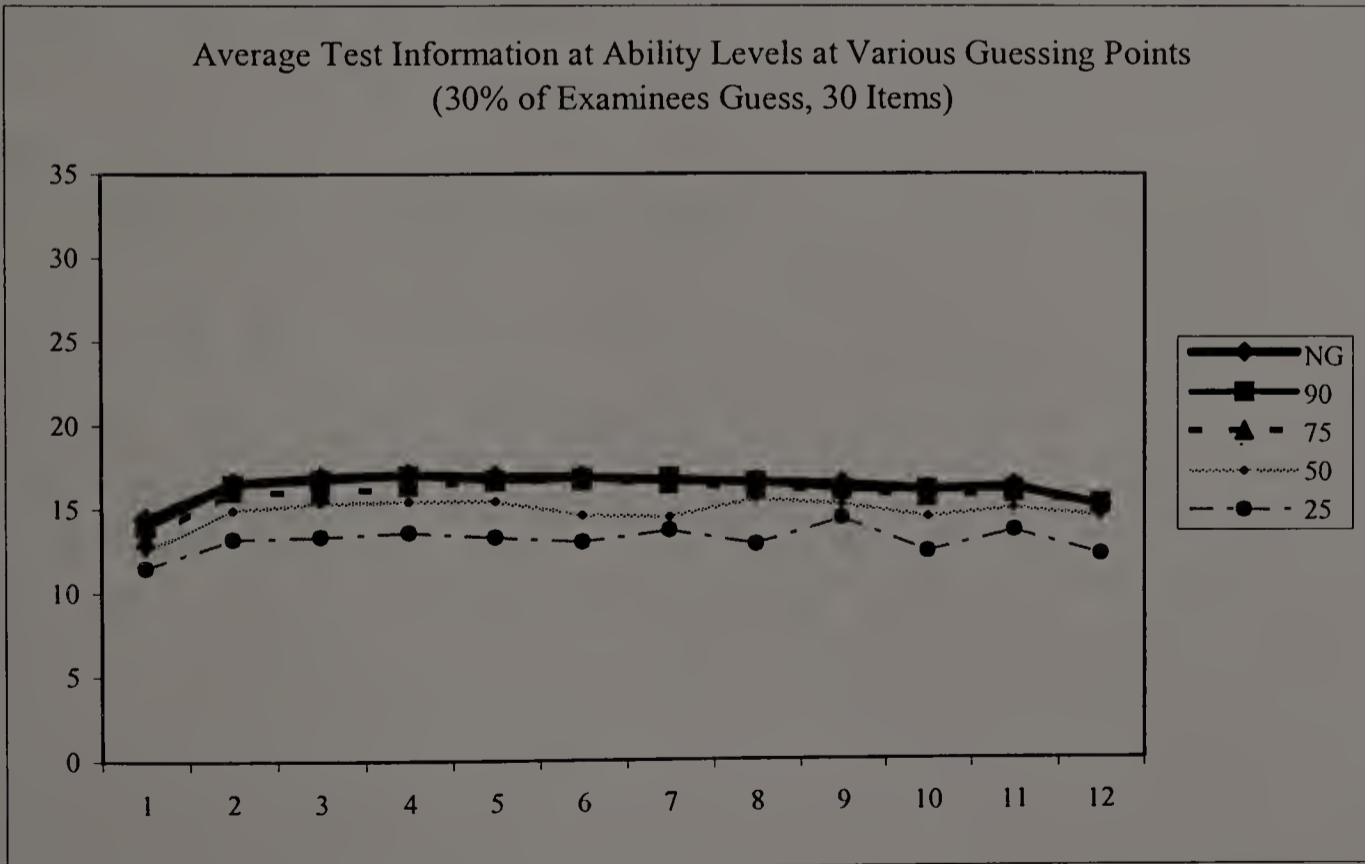
Scenario D: All flagged examinees begin guessing towards the end (after 90% of items)

Average Test Information at Ability Levels at Various Guessing Points  
 (Guessing Introduced for All vs. 30% of Examinees)

(50)



(51)





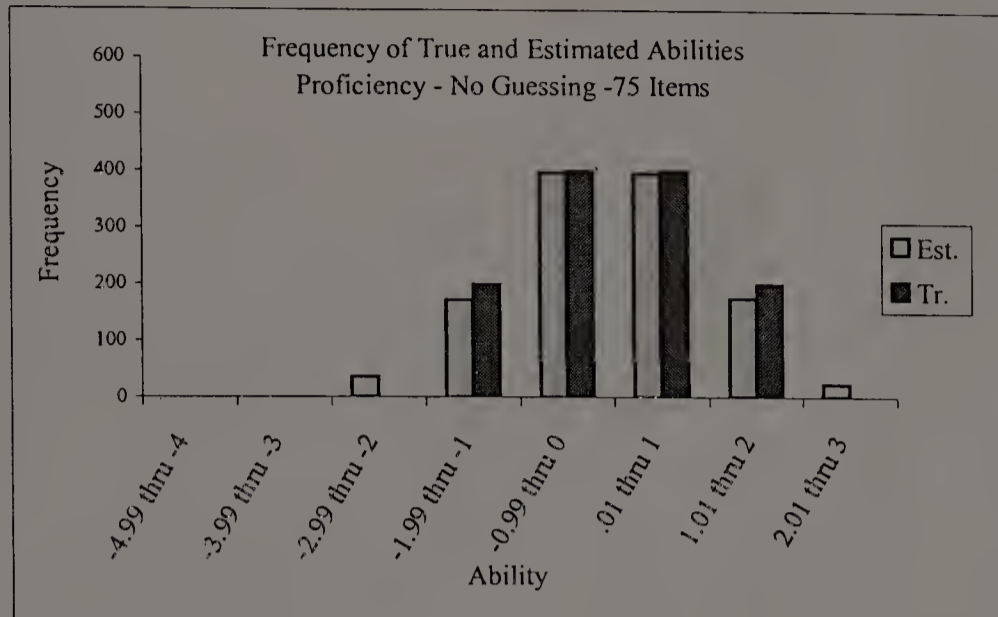
APPENDIX C

RESULTS USING AICPA ITEM PARAMETERS FOR AUDIT

## Distribution of Examinees in True and Estimated Ability Intervals

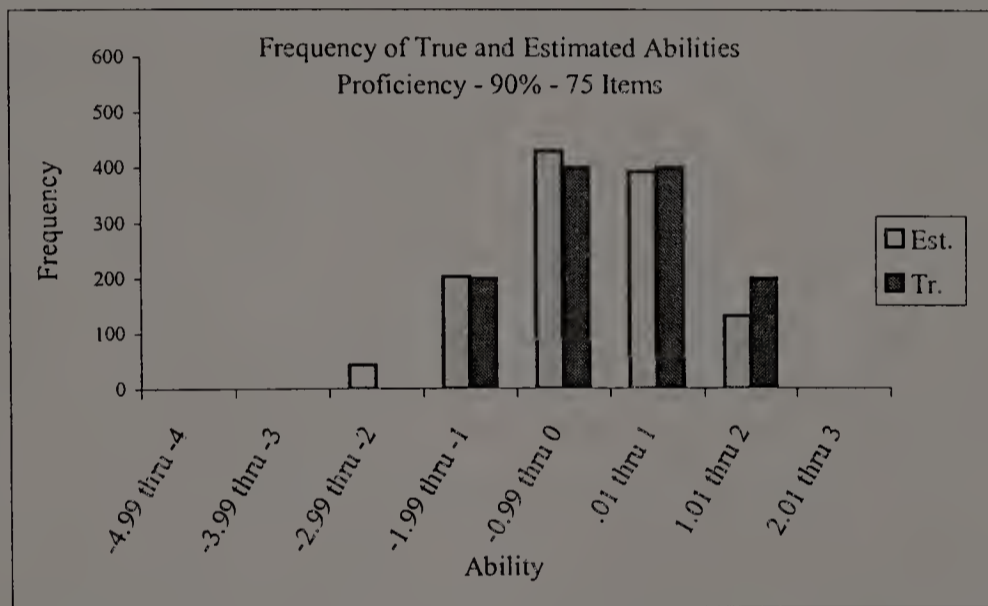
(52)

Ability	Est.	Tr.
-4.99 thru -4	0	0
-3.99 thru -3	0	0
-2.99 thru -2	36	0
-1.99 thru -1	172	200
-0.99 thru 0	398	400
.01 thru 1	396	400
1.01 thru 2	175	200
2.01 thru 3	23	0
Total	1200	0



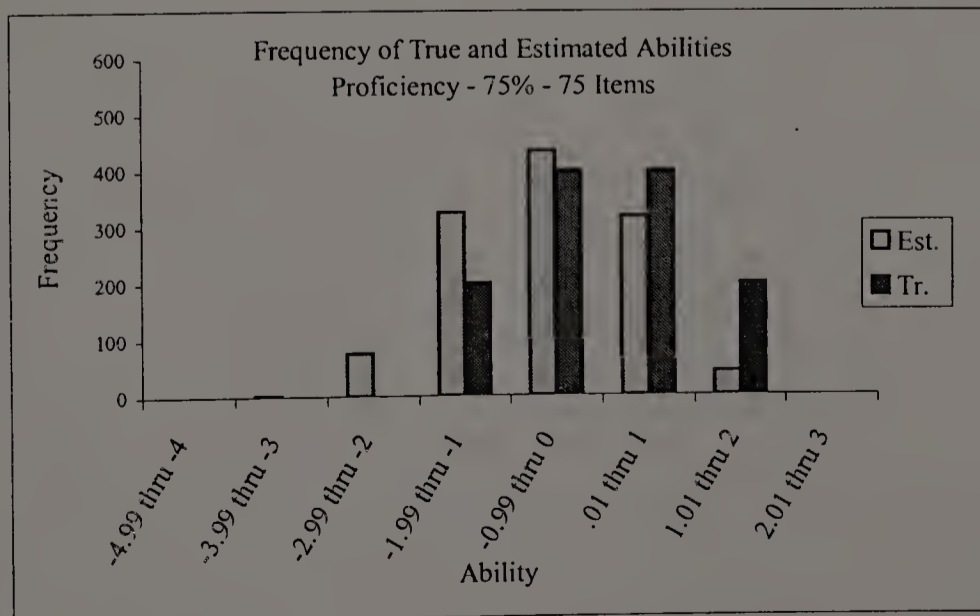
(53)

Ability	Est.	Tr.
-4.99 thru -4	0	0
-3.99 thru -3	0	0
-2.99 thru -2	43	0
-1.99 thru -1	203	200
-0.99 thru 0	430	400
.01 thru 1	393	400
1.01 thru 2	131	200
2.01 thru 3	0	0
Total	1200	0



(54)

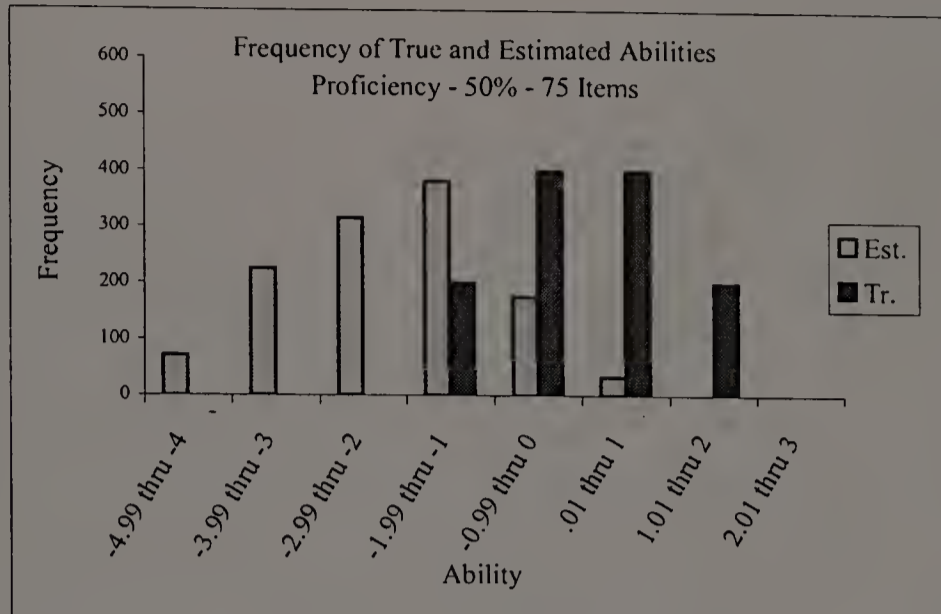
Ability	Est.	Tr.
-4.99 thru -4	0	0
-3.99 thru -3	2	0
-2.99 thru -2	76	0
-1.99 thru -1	326	200
-0.99 thru 0	435	400
.01 thru 1	318	400
1.01 thru 2	43	200
2.01 thru 3	0	0
Total	1200	0



## Distribution of Examinees in True and Estimated Ability Intervals

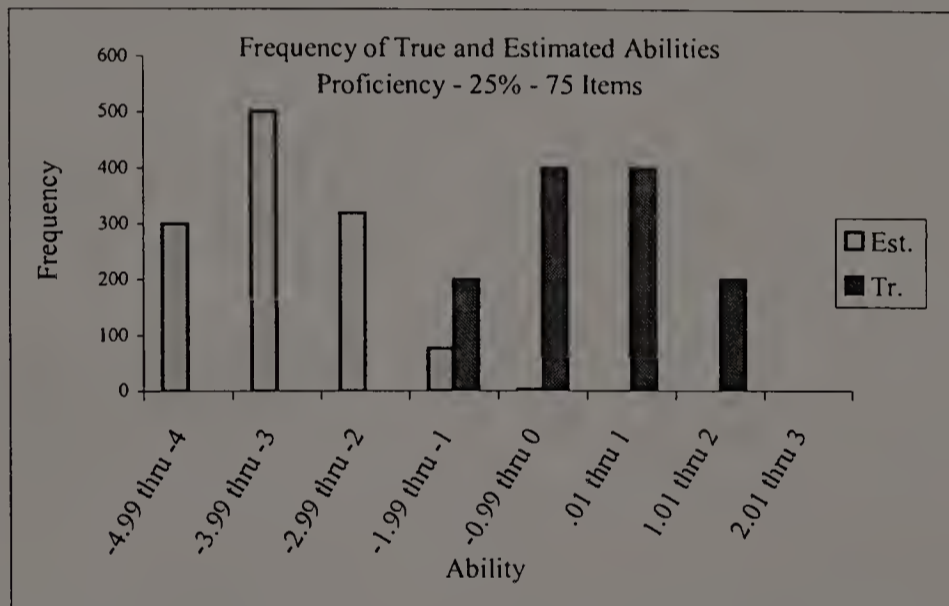
(55)

Ability	Est.	Tr.
-4.99 thru -4	71	0
-3.99 thru -3	225	0
-2.99 thru -2	315	0
-1.99 thru -1	381	200
-0.99 thru 0	175	400
.01 thru 1	33	400
1.01 thru 2	0	200
2.01 thru 3	0	0
Total	1200	0



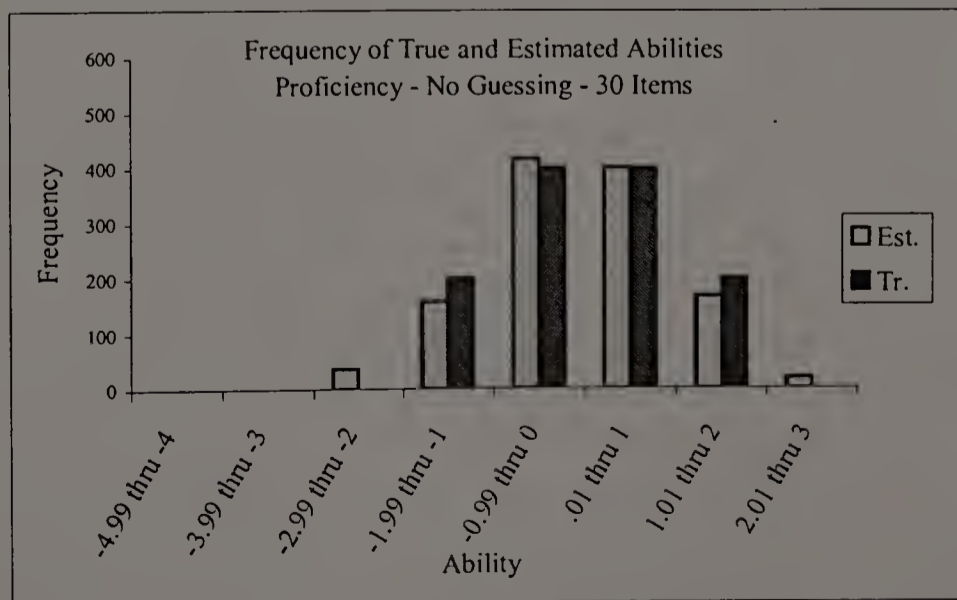
(56)

Ability	Est.	Tr.
-4.99 thru -4	300	0
-3.99 thru -3	501	0
-2.99 thru -2	320	0
-1.99 thru -1	76	200
-0.99 thru 0	3	400
.01 thru 1	0	400
1.01 thru 2	0	200
2.01 thru 3	0	0
Total	1200	0



(57)

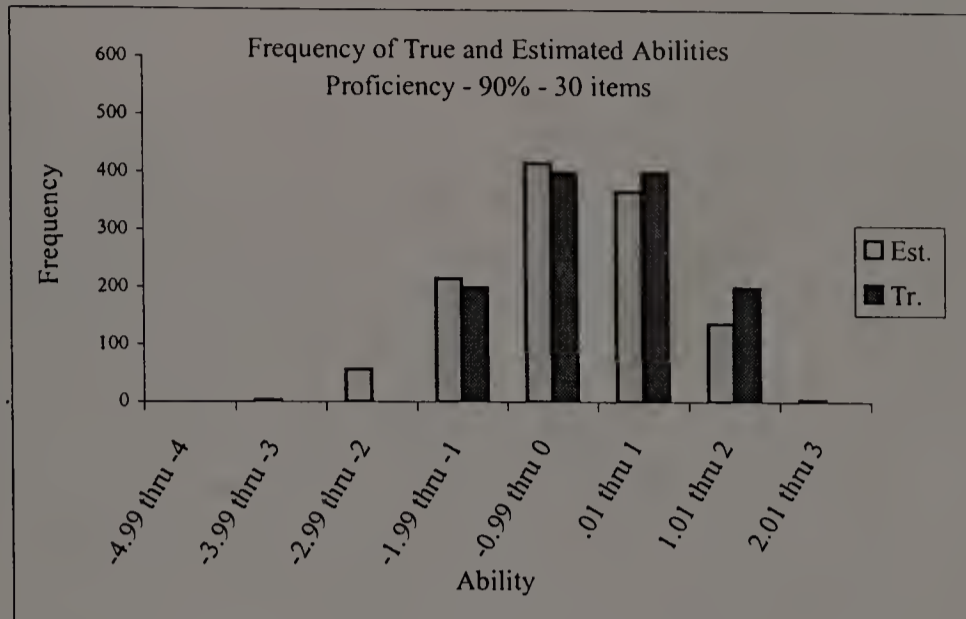
Ability	Est.	Tr.
-4.99 thru -4	0	0
-3.99 thru -3	0	0
-2.99 thru -2	37	0
-1.99 thru -1	158	200
-0.99 thru 0	417	400
.01 thru 1	401	400
1.01 thru 2	168	200
2.01 thru 3	19	0
Total	1200	0



## Distribution of Examinees in True and Estimated Ability Intervals

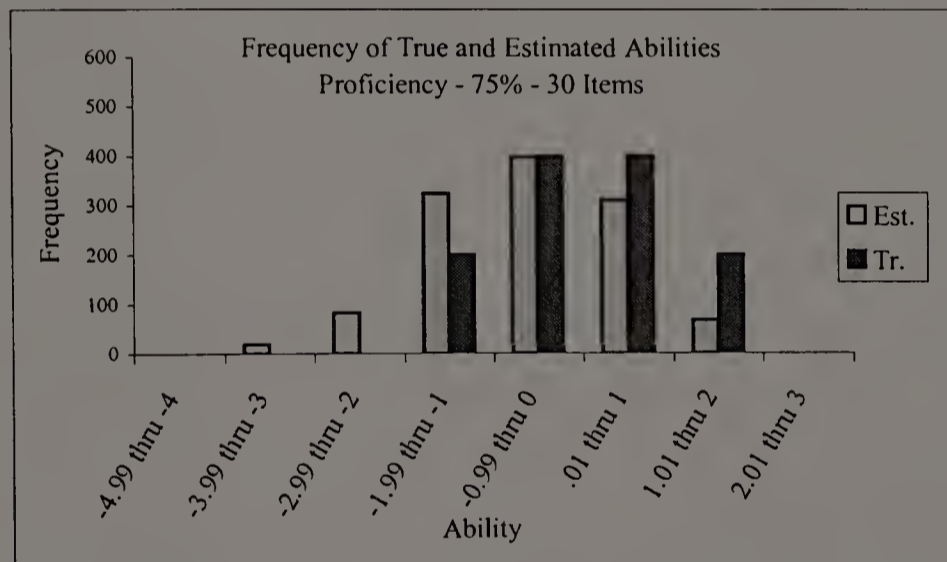
(58)

Ability	Est.	Tr.
-4.99 thru -4	0	0
-3.99 thru -3	4	0
-2.99 thru -2	57	0
-1.99 thru -1	215	200
-0.99 thru 0	416	400
.01 thru 1	367	400
1.01 thru 2	137	200
2.01 thru 3	4	0
Total	1200	0



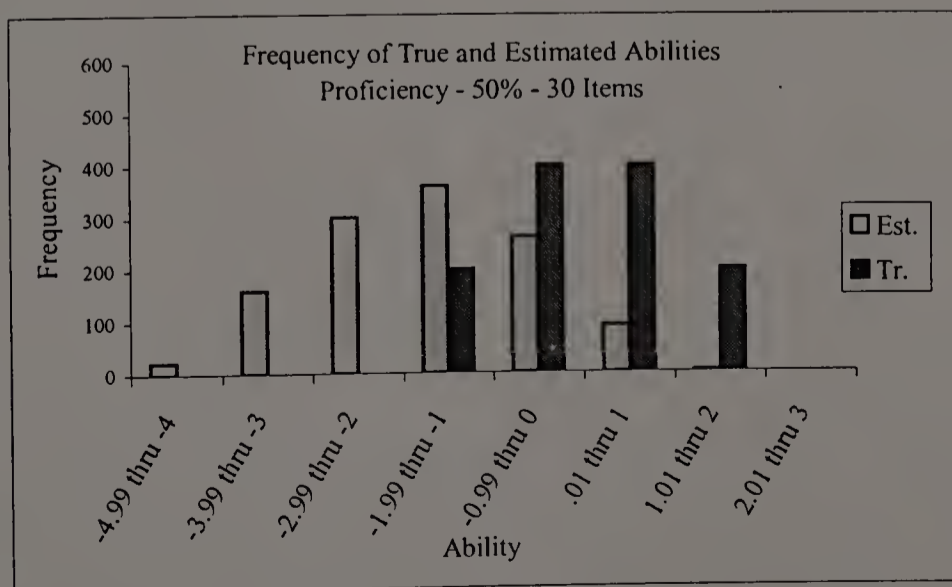
(59)

Ability	Est.	Tr.
-4.99 thru -4	0	0
-3.99 thru -3	19	0
-2.99 thru -2	83	0
-1.99 thru -1	324	200
-0.99 thru 0	397	400
.01 thru 1	310	400
1.01 thru 2	67	200
2.01 thru 3	0	0
Total	1200	0



(60)

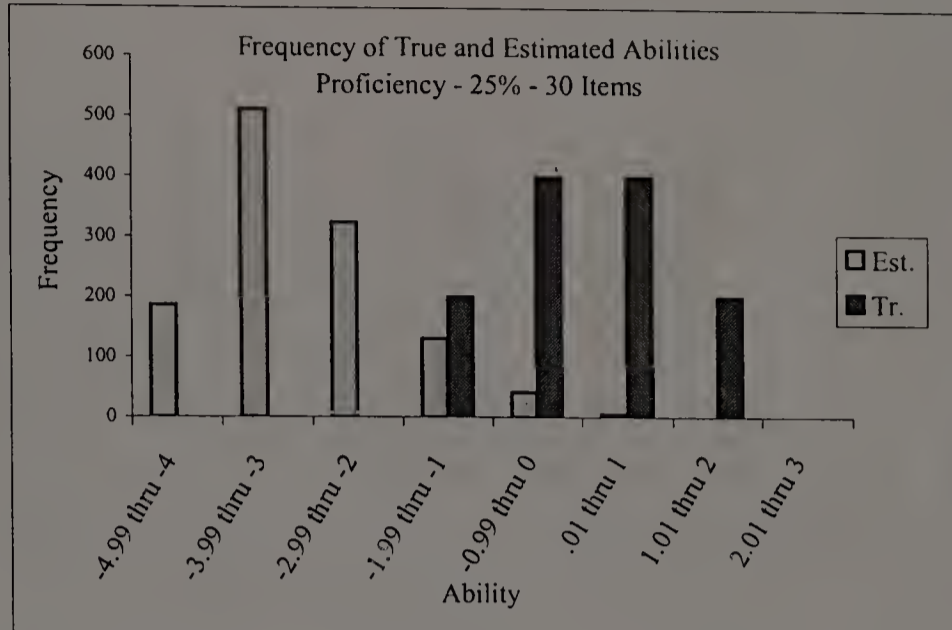
Ability	Est.	Tr.
-4.99 thru -4	23	0
-3.99 thru -3	160	0
-2.99 thru -2	301	0
-1.99 thru -1	361	200
-0.99 thru 0	262	400
.01 thru 1	91	400
1.01 thru 2	2	200
2.01 thru 3	0	0
Total	1200	0



## Distribution of Examinees in True and Estimated Ability Intervals

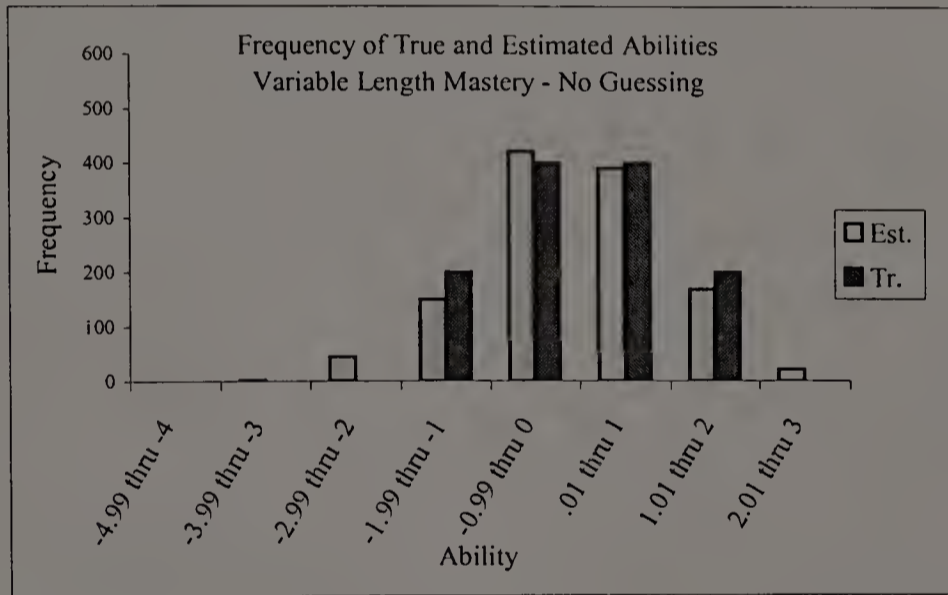
(61)

Ability	Est.	Tr.
-4.99 thru -4	186	0
-3.99 thru -3	511	0
-2.99 thru -2	324	0
-1.99 thru -1	131	200
-0.99 thru 0	42	400
.01 thru 1	6	400
1.01 thru 2	0	200
2.01 thru 3	0	0
Total	1200	0



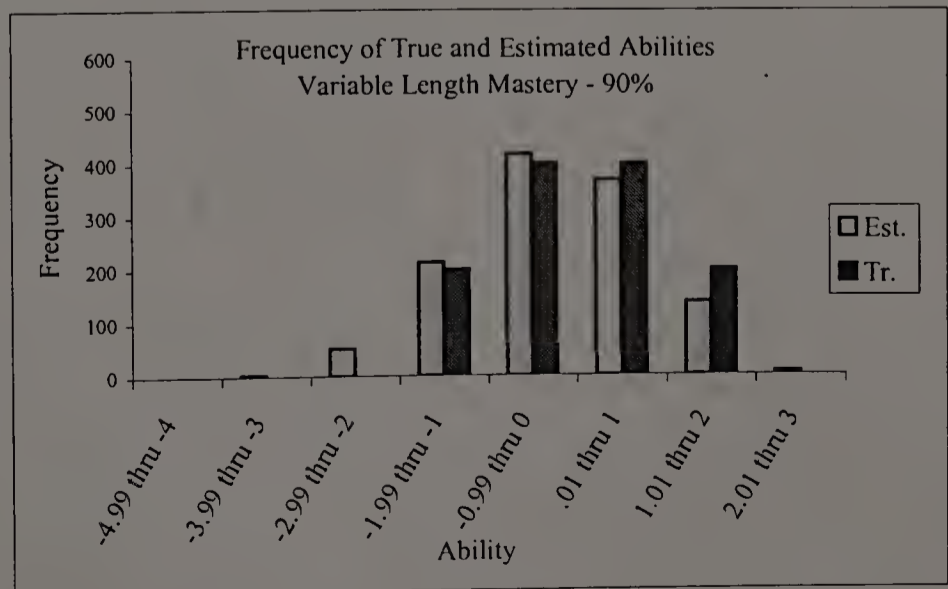
(62)

Ability	Est.	Tr.
-4.99 thru -4	0	0
-3.99 thru -3	2	0
-2.99 thru -2	45	0
-1.99 thru -1	150	200
-0.99 thru 0	422	400
.01 thru 1	391	400
1.01 thru 2	168	200
2.01 thru 3	22	0
Total	1200	0



(63)

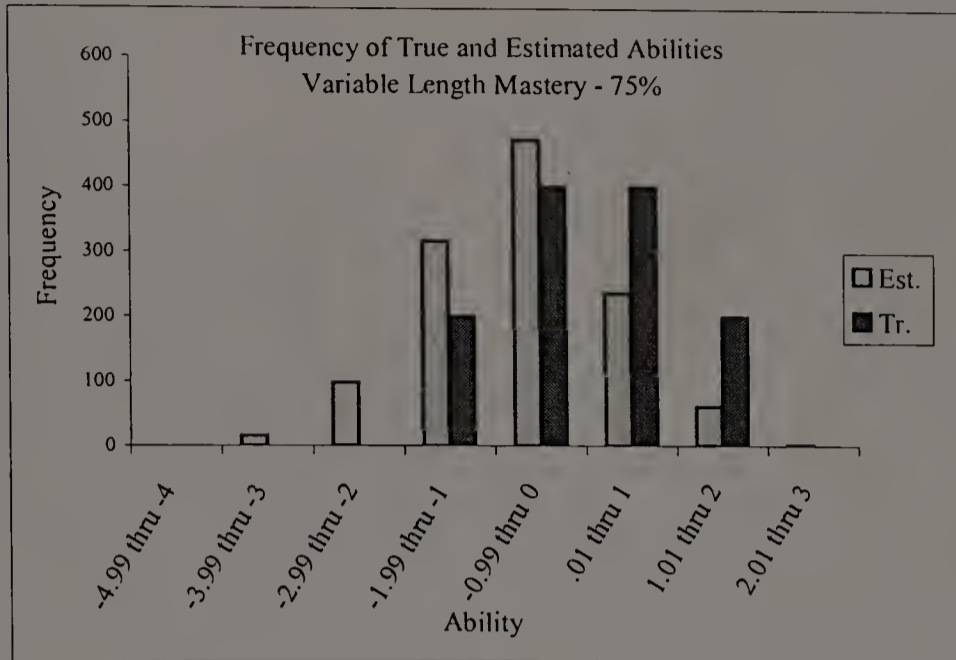
Ability	Est.	Tr.
-4.99 thru -4	0	0
-3.99 thru -3	3	0
-2.99 thru -2	52	0
-1.99 thru -1	214	200
-0.99 thru 0	418	400
.01 thru 1	369	400
1.01 thru 2	138	200
2.01 thru 3	6	0
Total	1200	0



## Distribution of Examinees in True and Estimated Ability Intervals

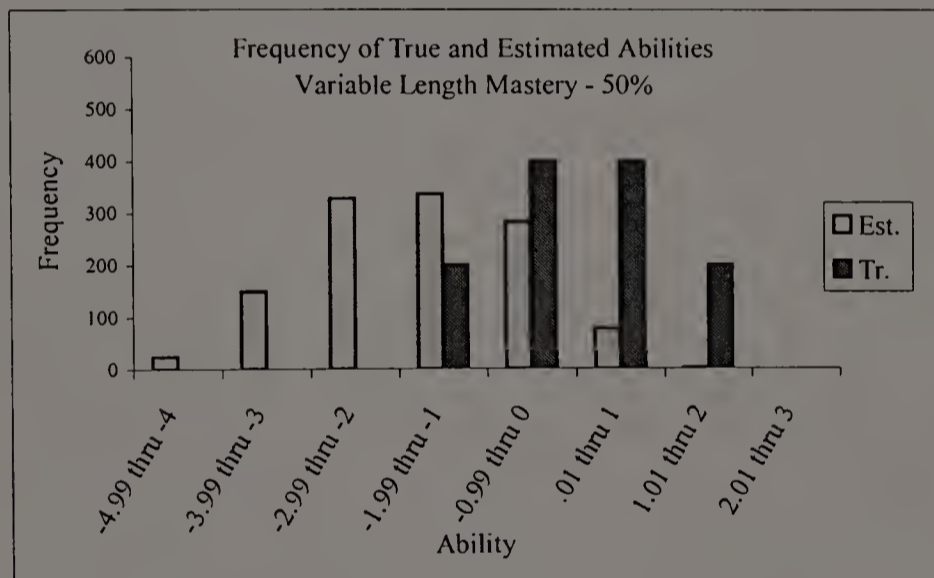
(64)

Ability	Est.	Tr.
-4.99 thru -4	0	0
-3.99 thru -3	16	0
-2.99 thru -2	98	0
-1.99 thru -1	316	200
-0.99 thru 0	472	400
.01 thru 1	236	400
1.01 thru 2	61	200
2.01 thru 3	1	0
Total	1200	0



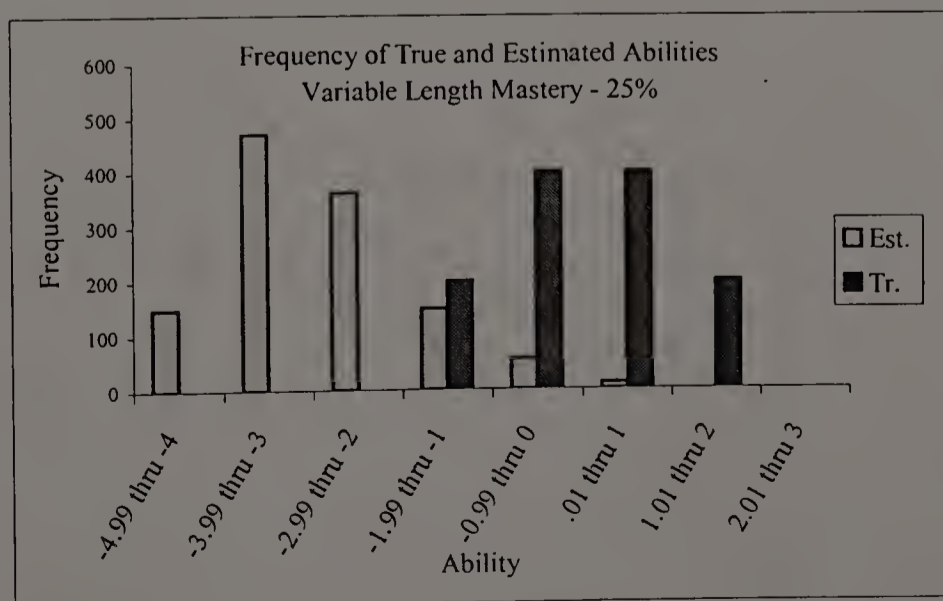
(65)

Ability	Est.	Tr.
-4.99 thru -4	24	0
-3.99 thru -3	149	0
-2.99 thru -2	328	0
-1.99 thru -1	337	200
-0.99 thru 0	283	400
.01 thru 1	77	400
1.01 thru 2	2	200
2.01 thru 3	0	0
Total	1200	0



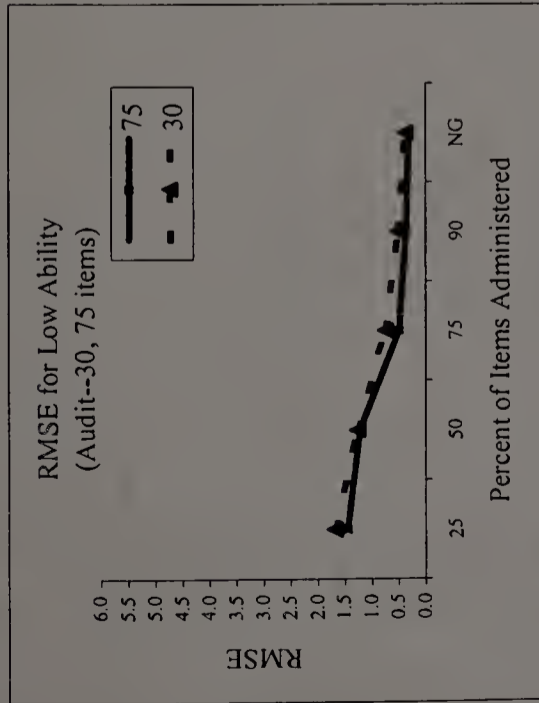
(66)

Ability	Est.	Tr.
-4.99 thru -4	149	0
-3.99 thru -3	471	0
-2.99 thru -2	363	0
-1.99 thru -1	149	200
-0.99 thru 0	56	400
.01 thru 1	12	400
1.01 thru 2	0	200
2.01 thru 3	0	0
Total	1200	0

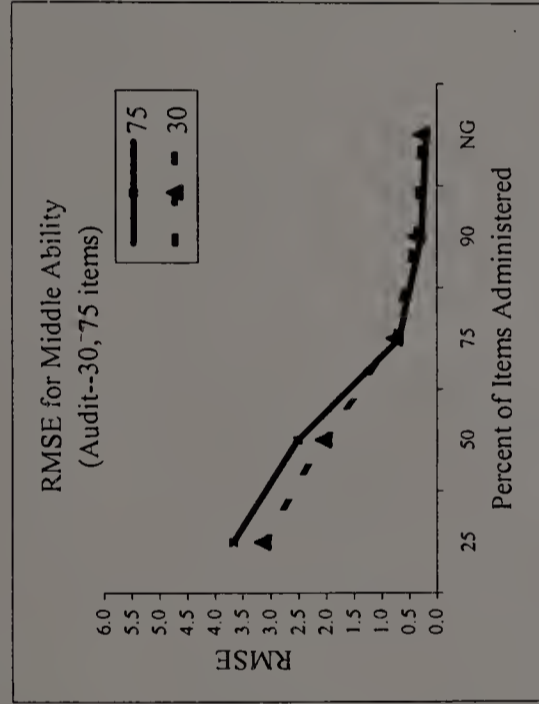


RMSE of Estimates around True Ability for 5 Guessing Scenarios  
 (Performance Testing with AICPA Parameters for 30 and 75 Items on AUDIT)

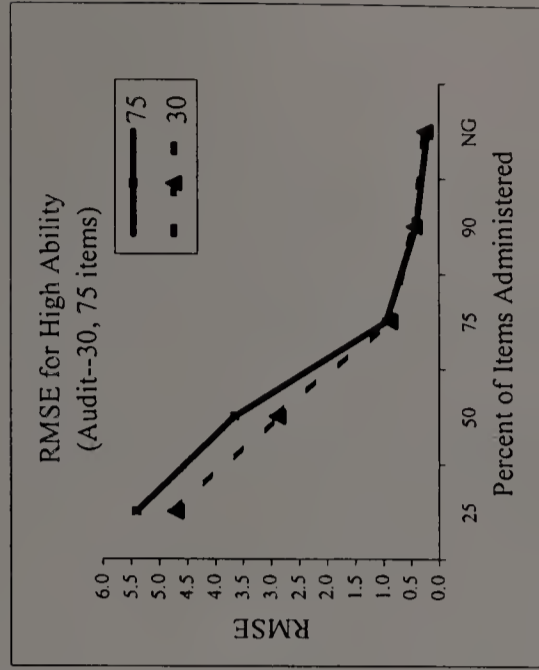
(67)



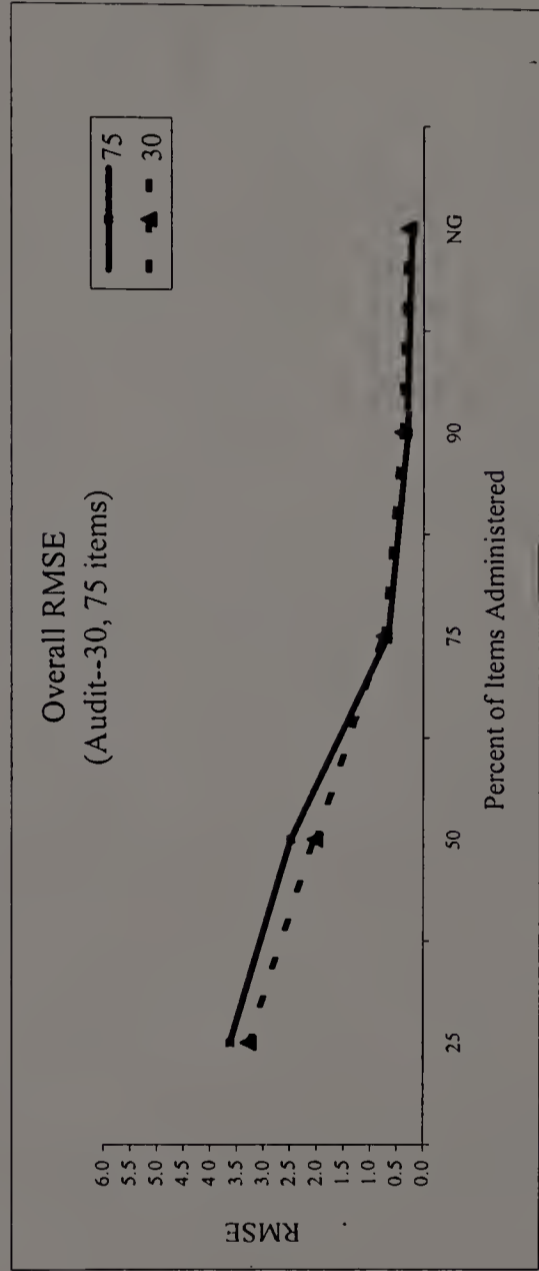
(68)



(69)

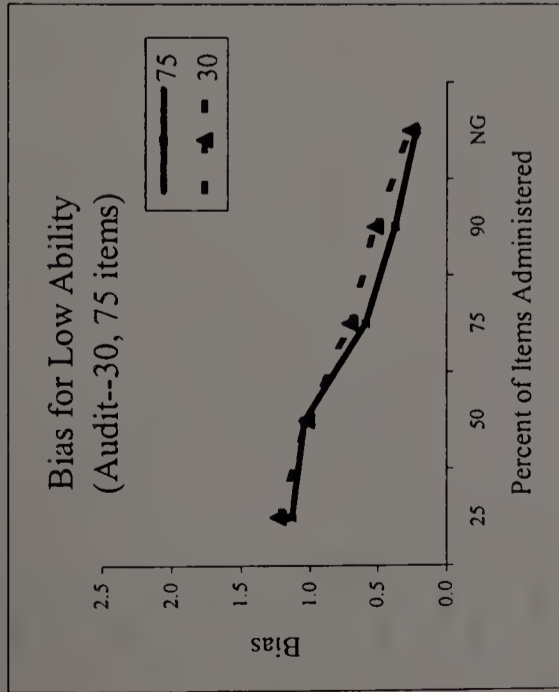


(70)

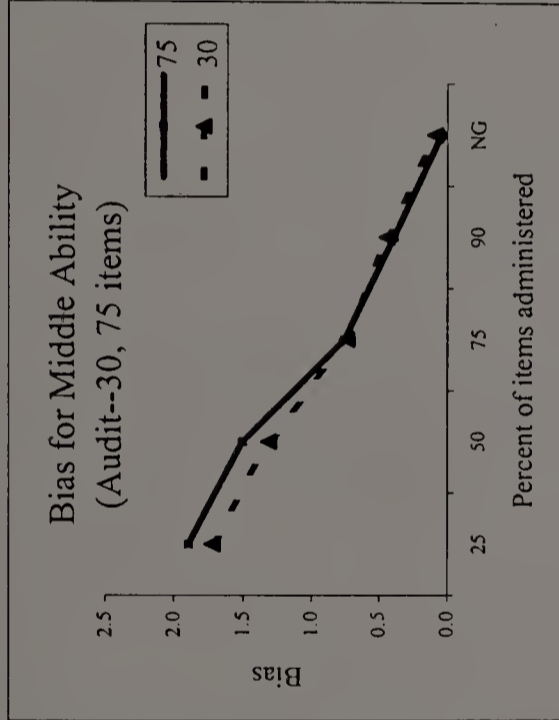


Bias in Estimates around True ability for 5 Guessing Scenarios  
 (Performance Testing with AICPA Parameters for 30 and 75 Items on AUDIT)

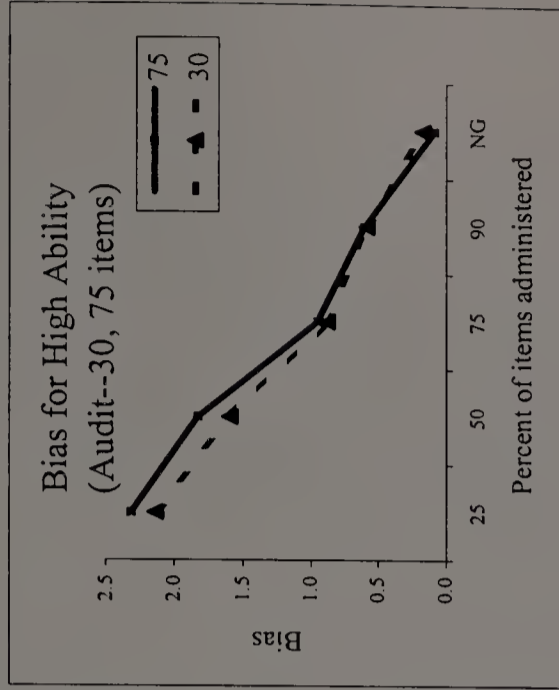
(71)



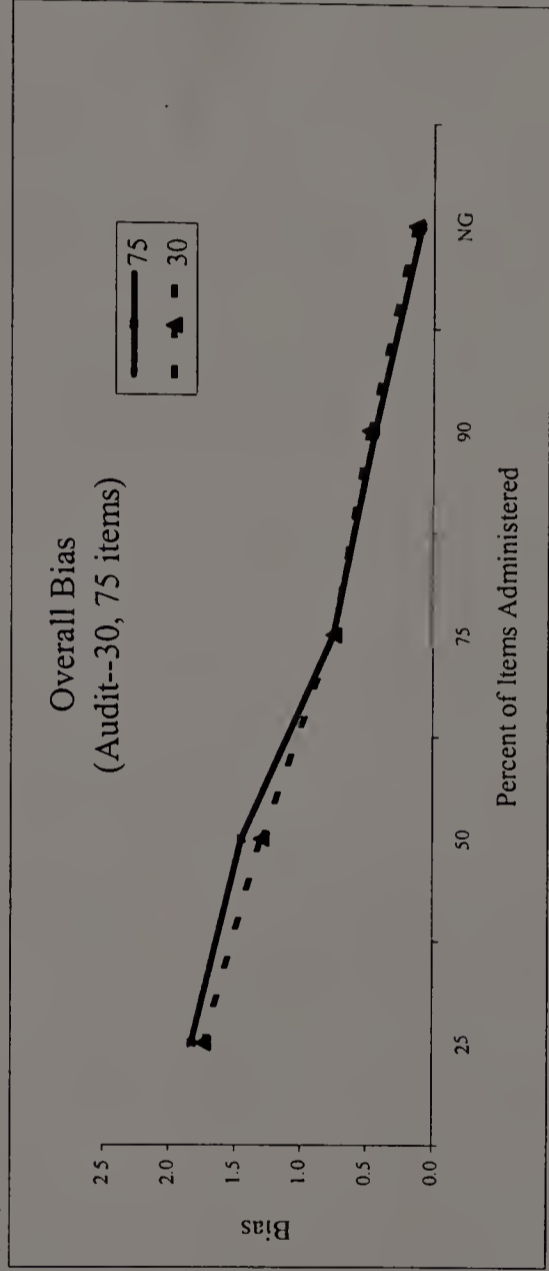
(72)



(73)



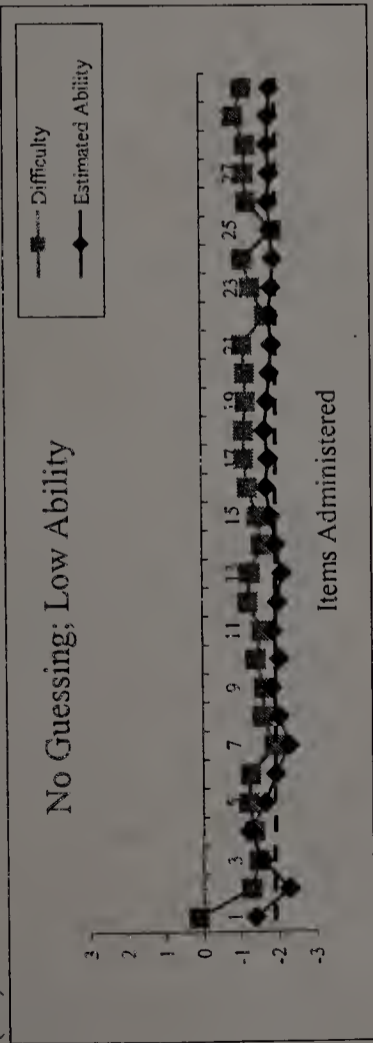
(74)



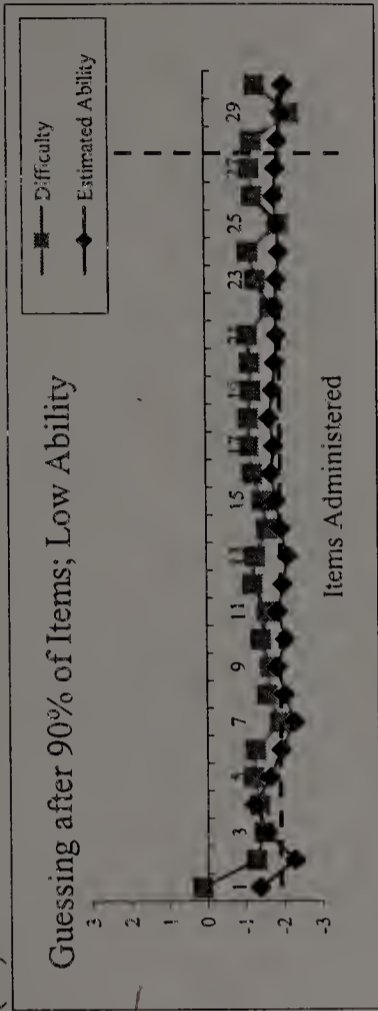


CAT Administration for a Low Ability Examinee who Guesses at a Certain Point in the Test (AUDIT--30 Items)

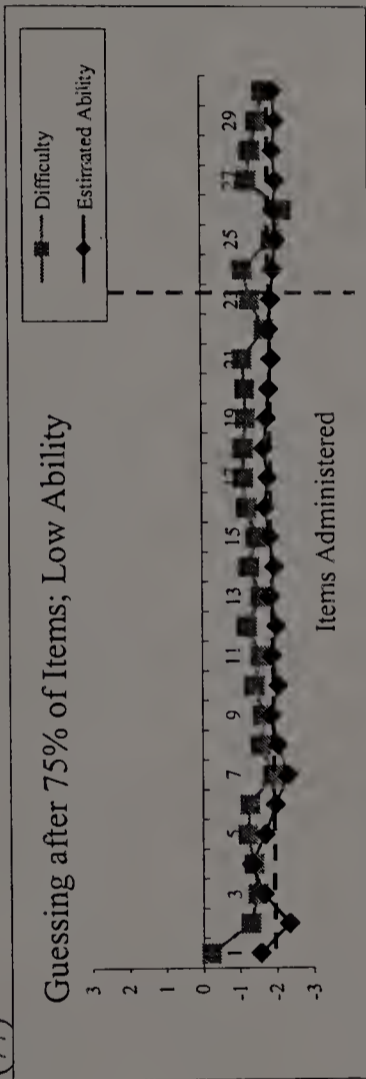
(75)



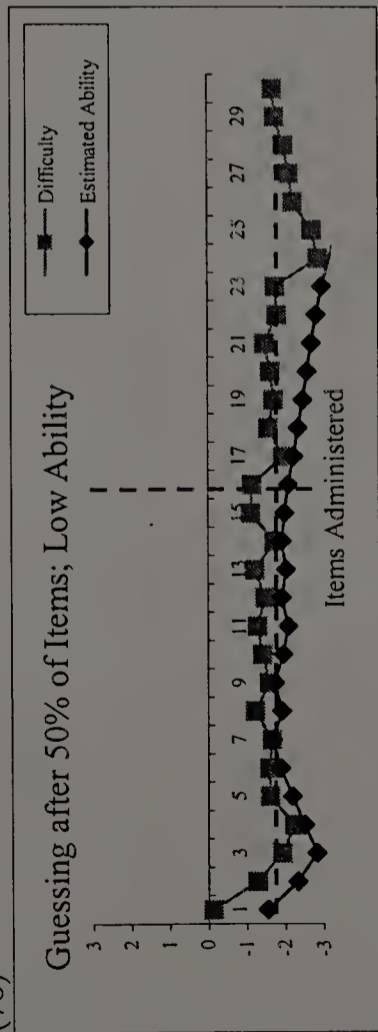
(76)



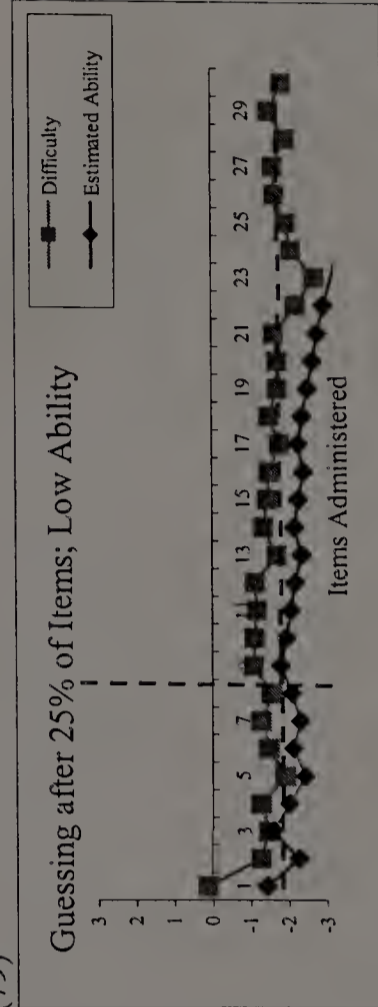
(77)



(78)

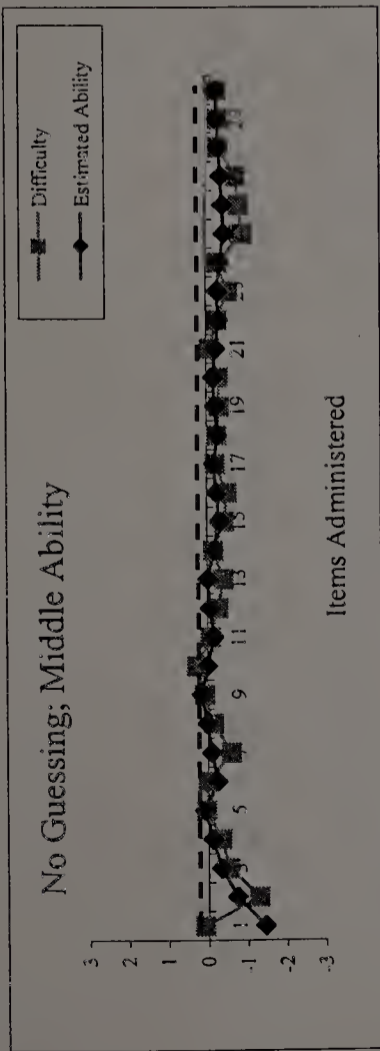


(79)

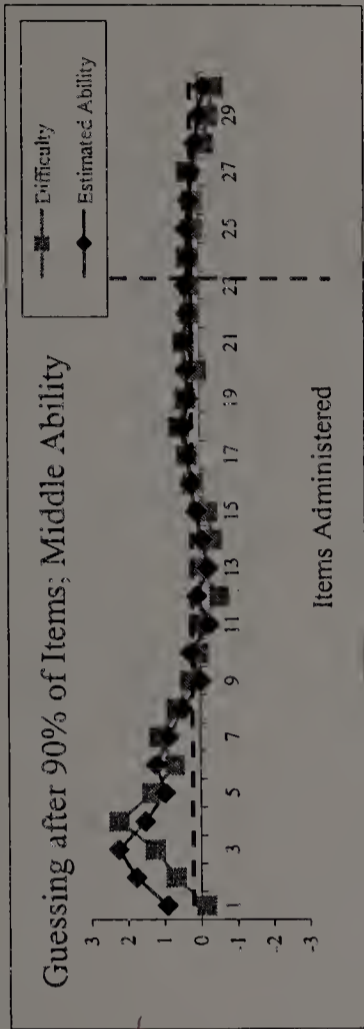


CAT Administration for a Middle Ability Examinee who Guesses at a Certain Point in the Test (AUDIT--30 Items)

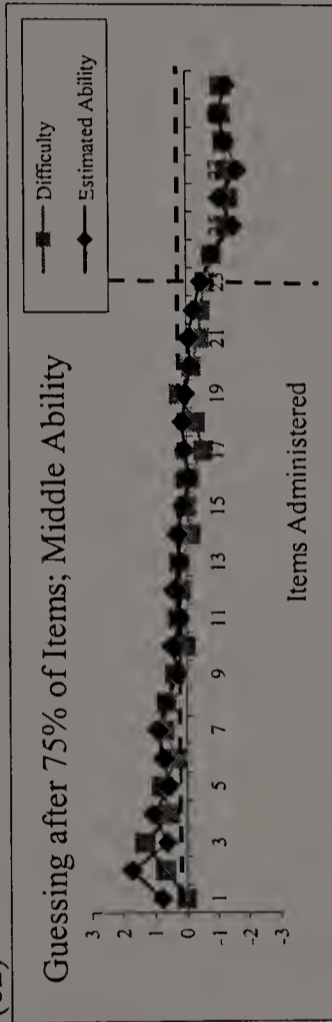
(80)



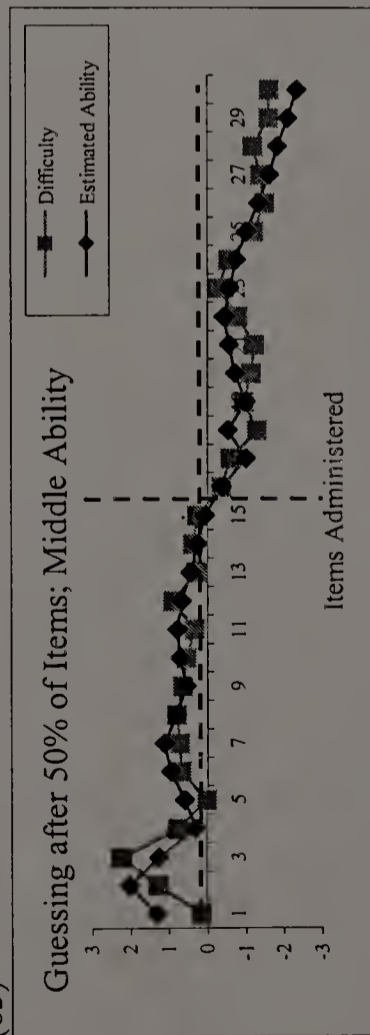
(81)



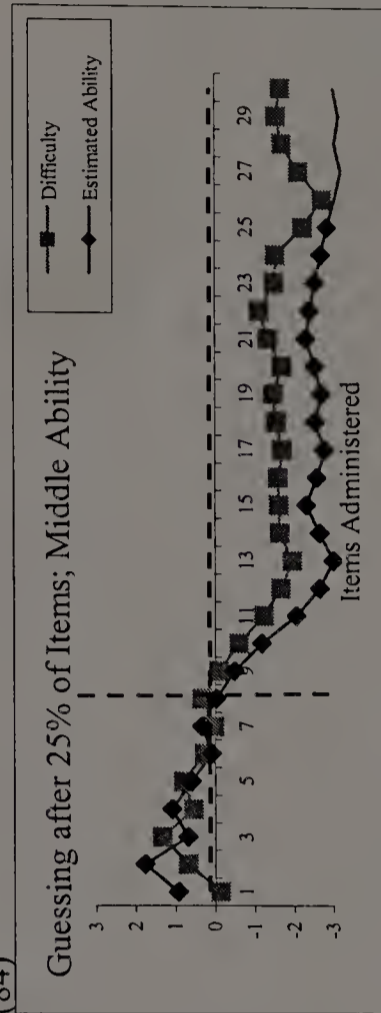
(82)



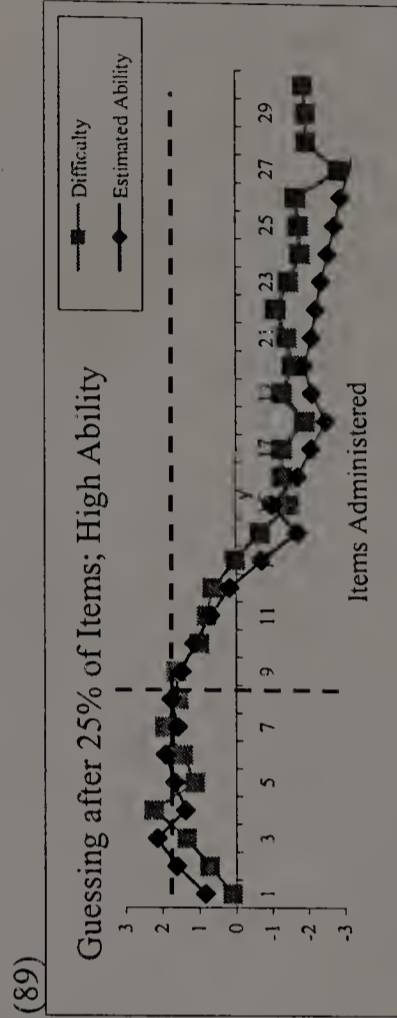
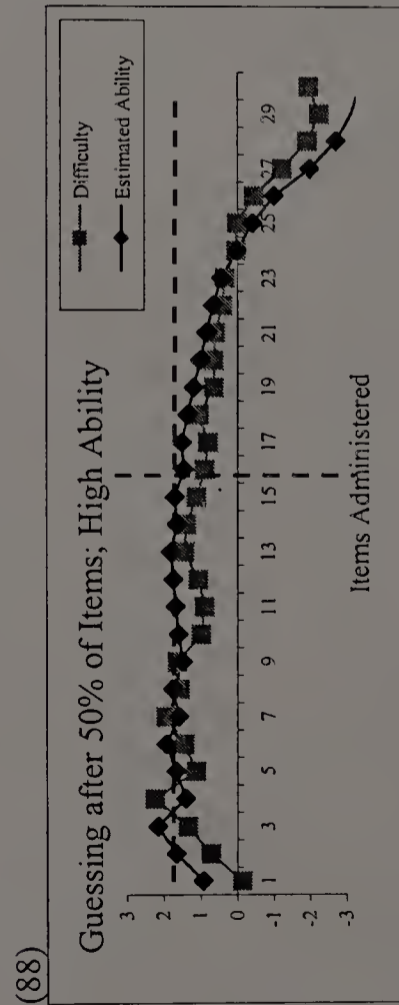
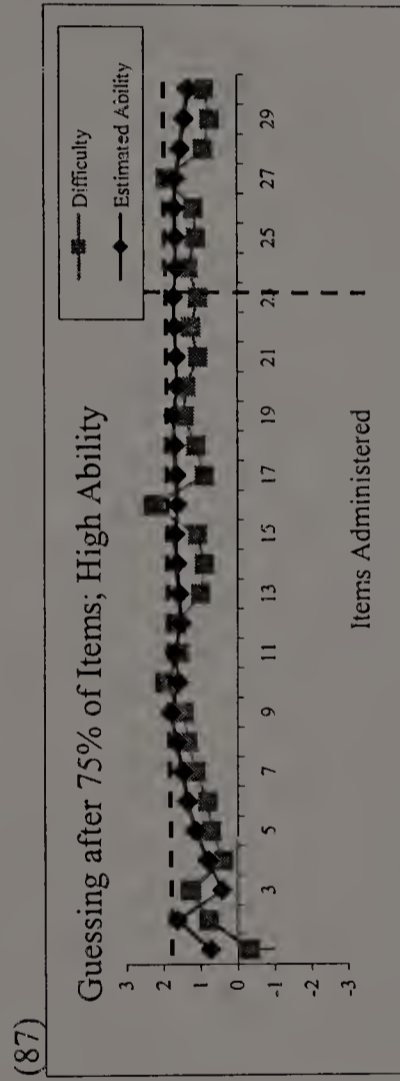
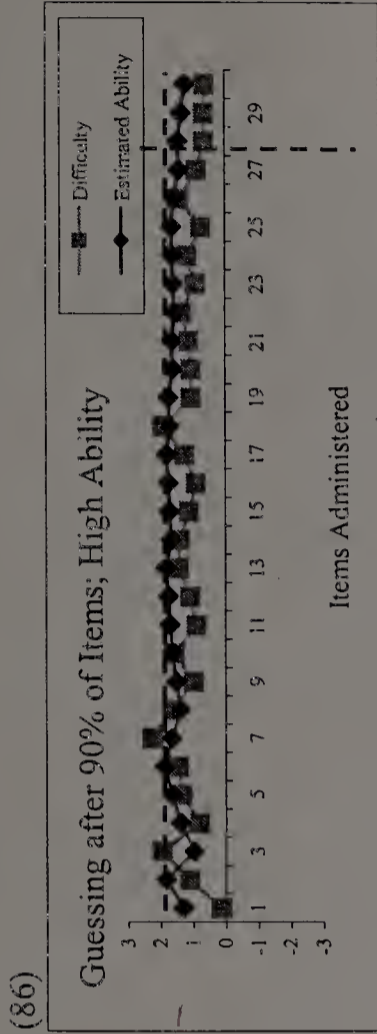
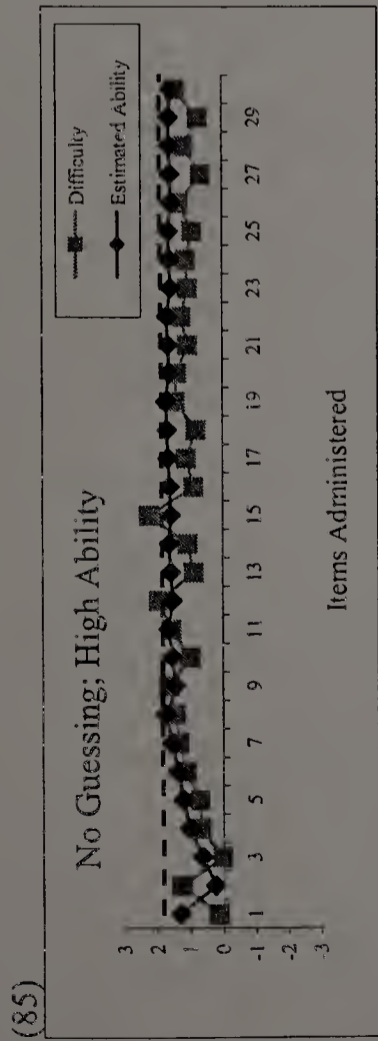
(83)



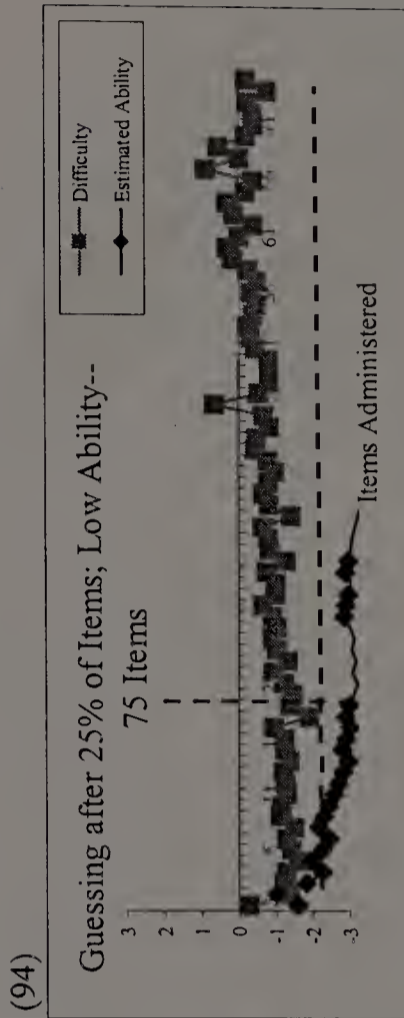
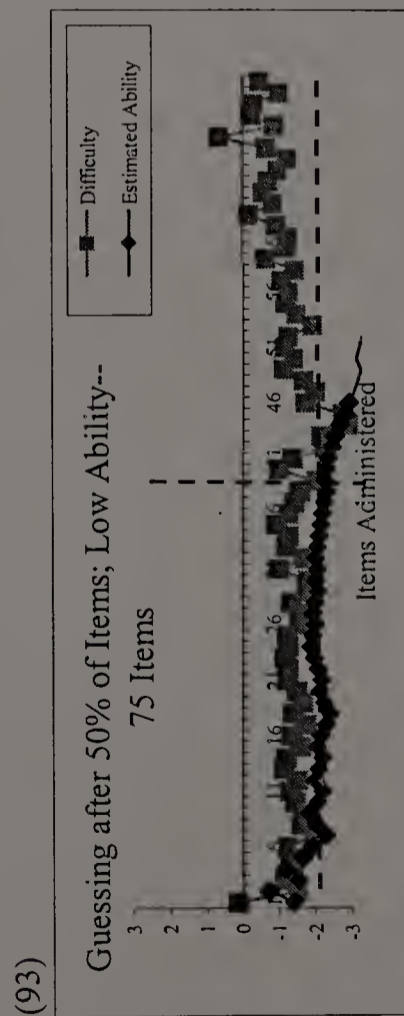
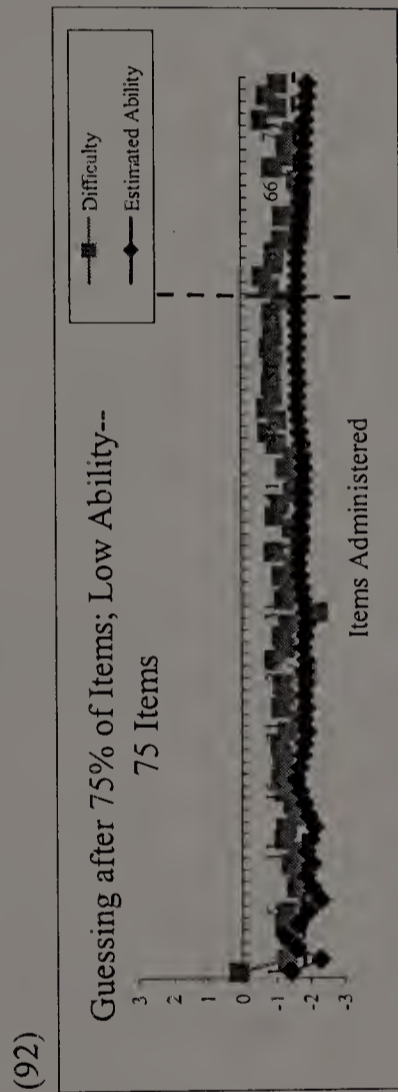
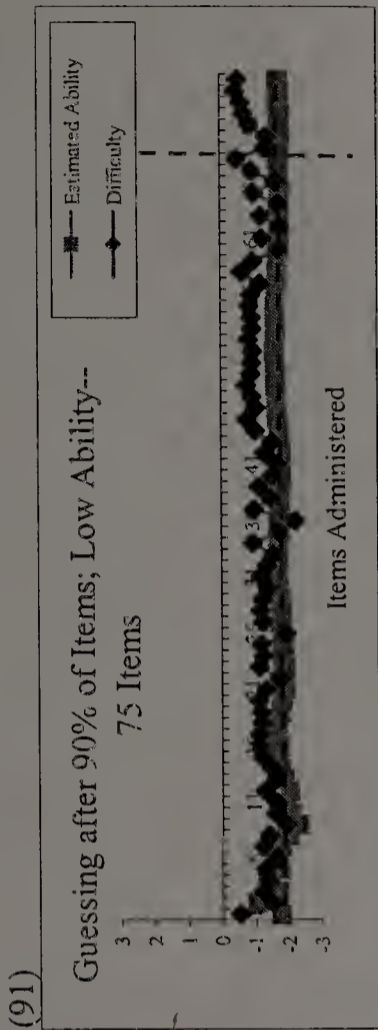
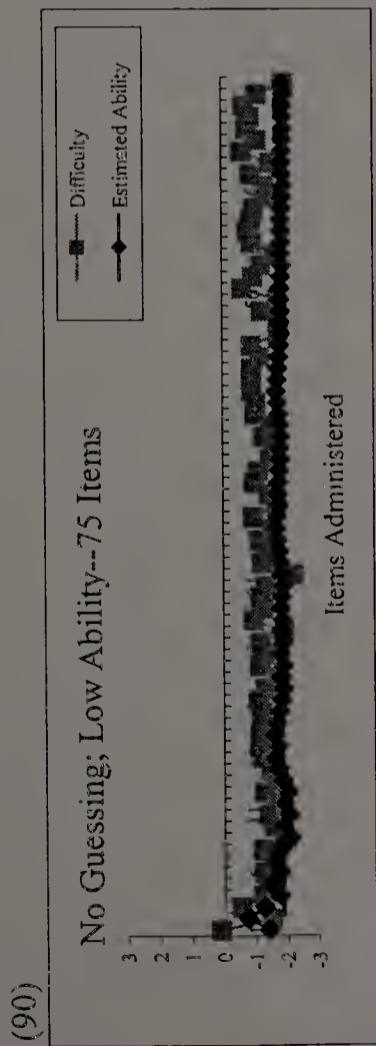
(84)



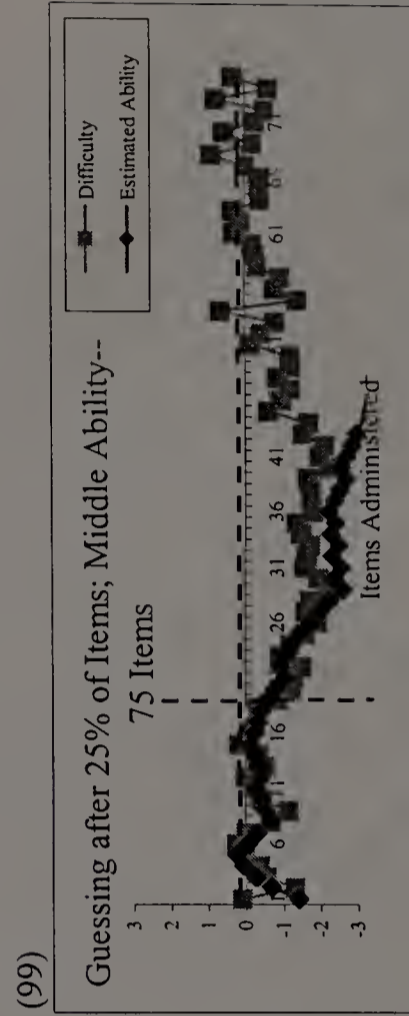
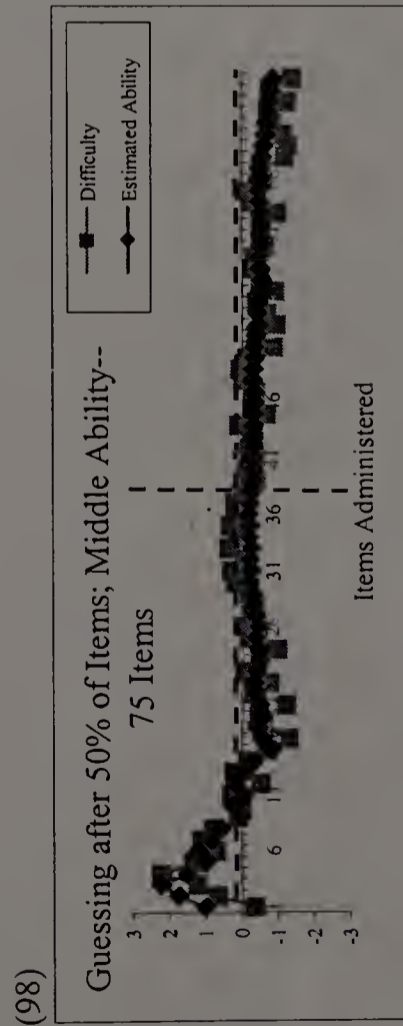
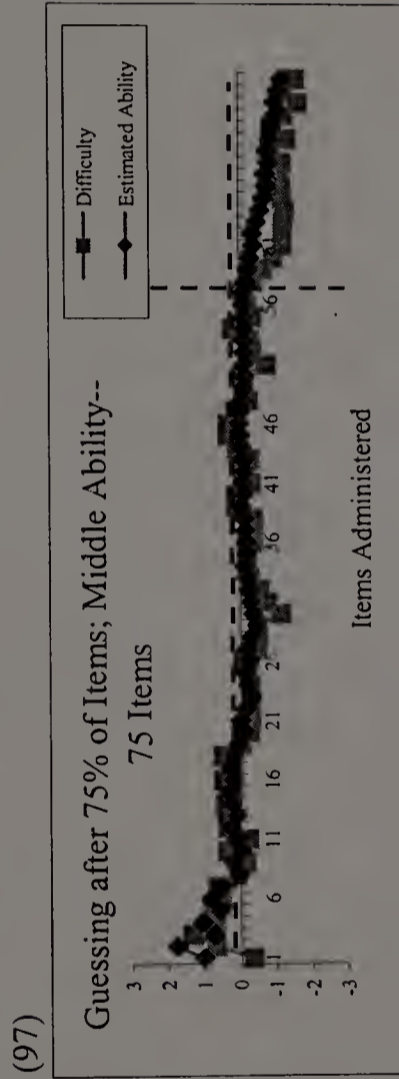
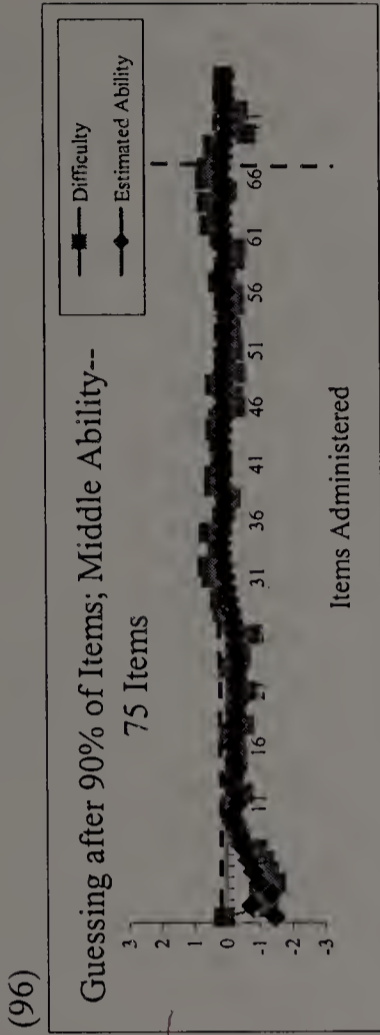
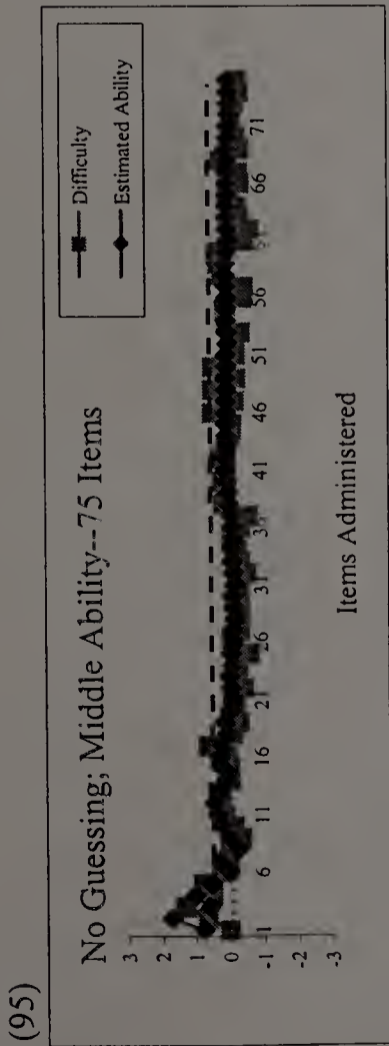
CAT Administration for a High Ability Examinee who Guesses at a Certain Point in the Test (AUDIT--30 Items)



CAT Administration for a Low Ability Examinee who Guesses at a Certain Point in the Test (AUDIT--75 Items)

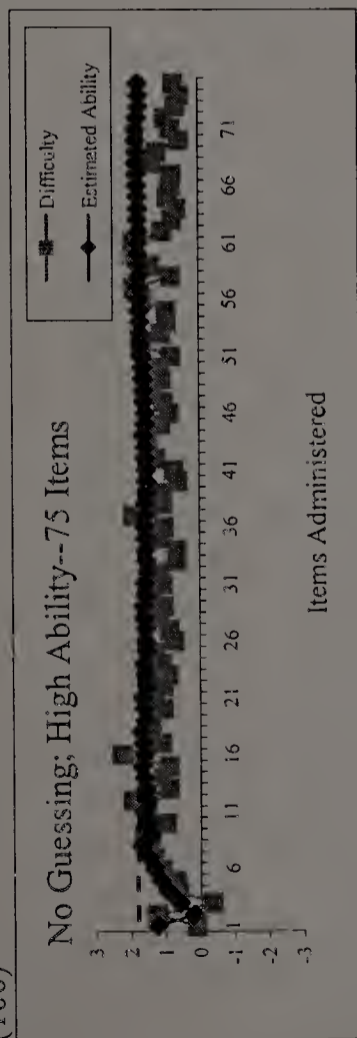


CAT Administration for a Middle Ability Examinee who Guesses at a Certain Point in the Test (AUDIT--75 Items)

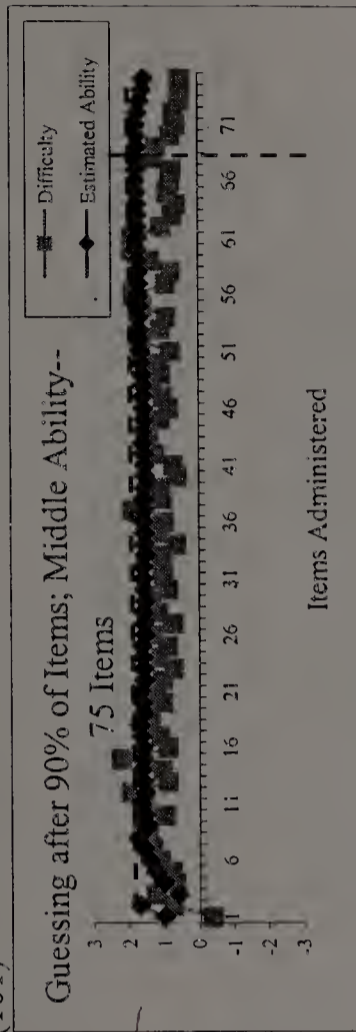


CAT Administration for a High Ability Examinee who Guesses at a Certain Point in the Test (AUDIT--75 Items)

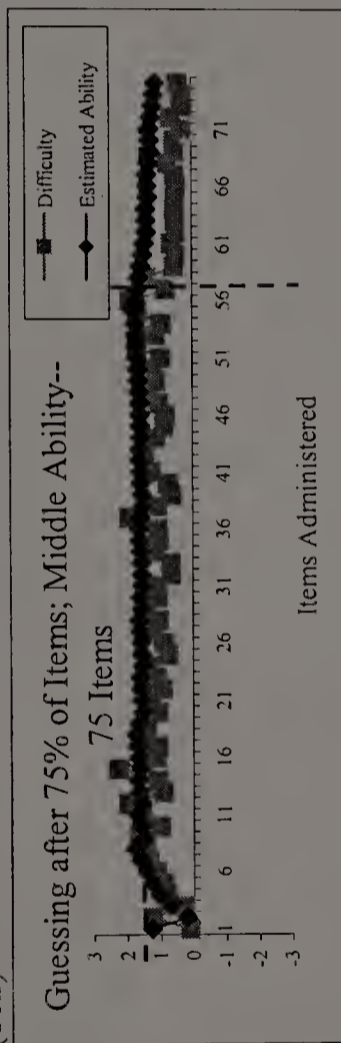
(100)



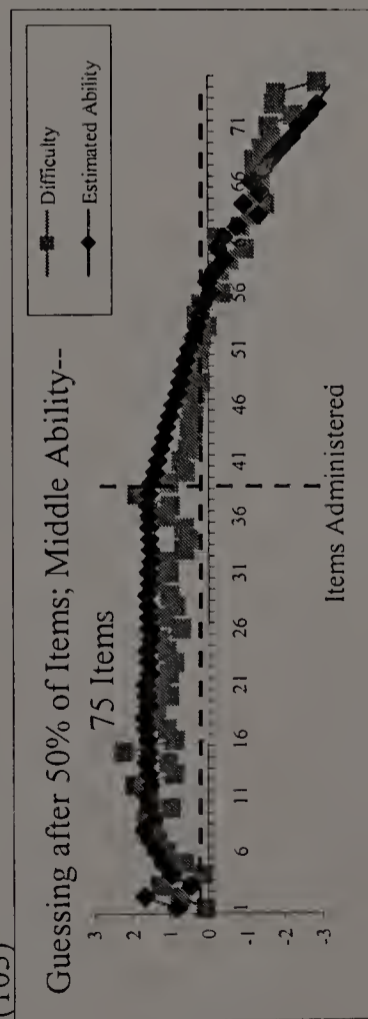
(101)



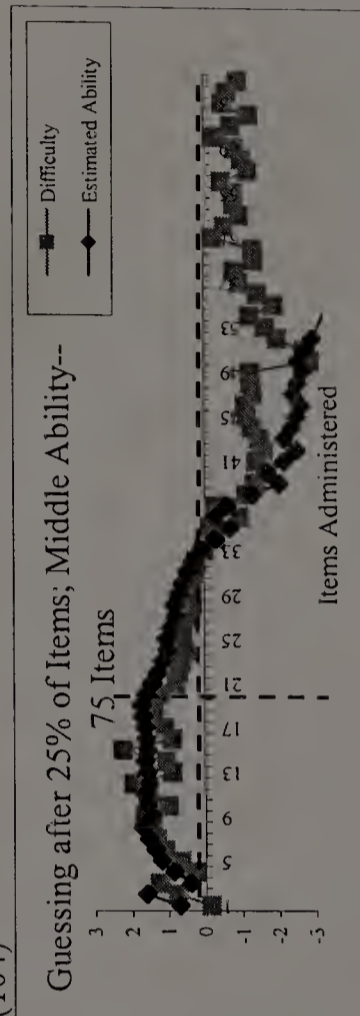
(102)



(103)

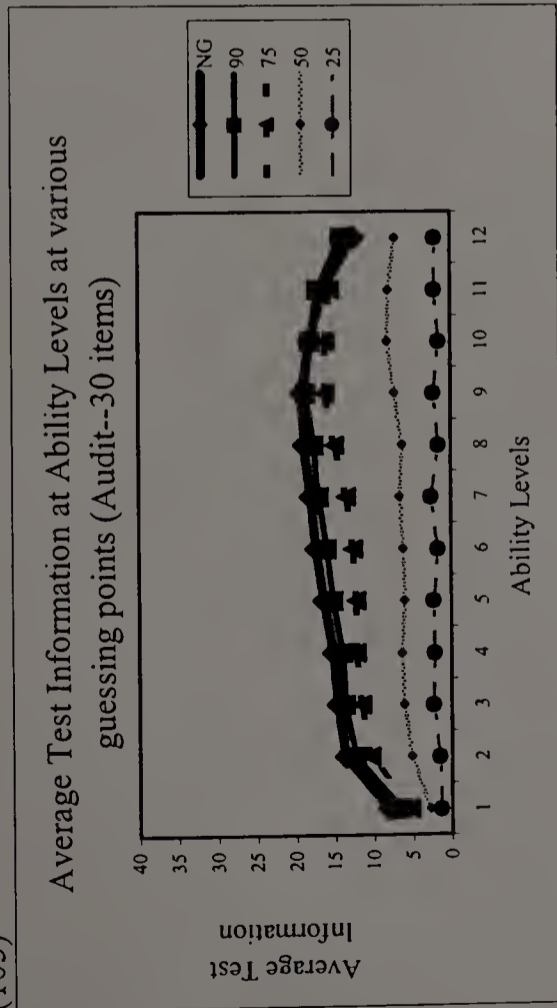


(104)

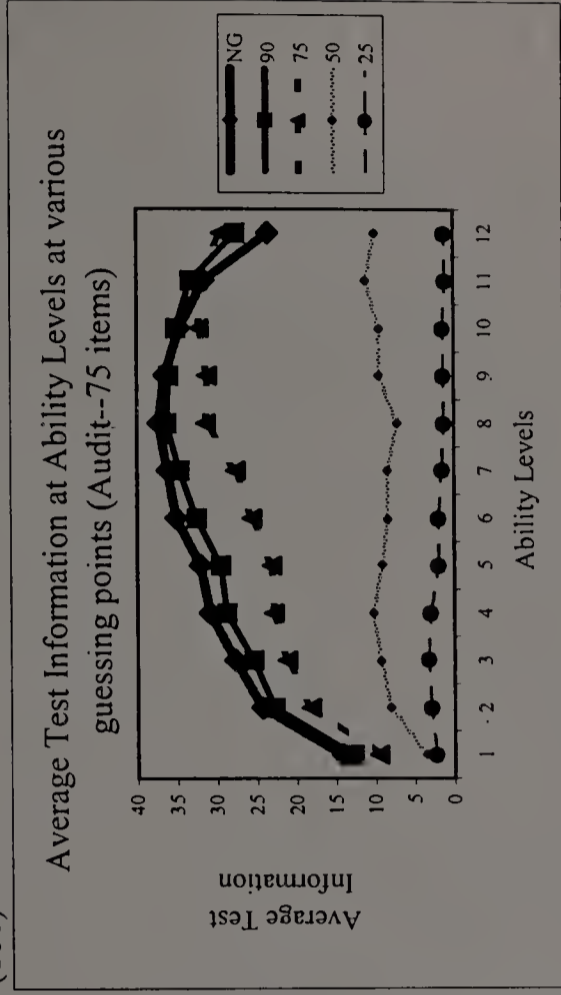


Average Test and Pool Information at each Ability Level for 5 Guessing Scenarios  
 (Performance Testing with AICPA Parameters for 30 and 75 Items on AUDIT)

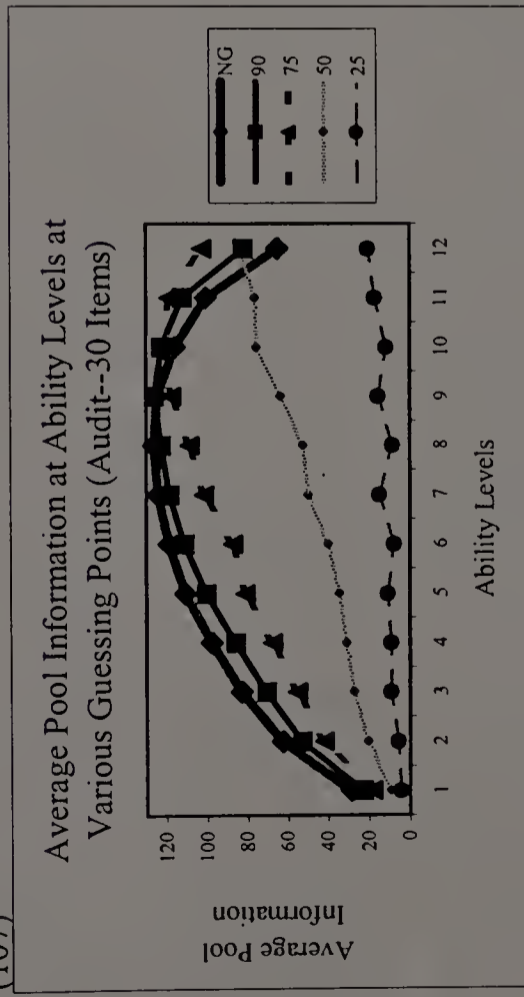
(105)



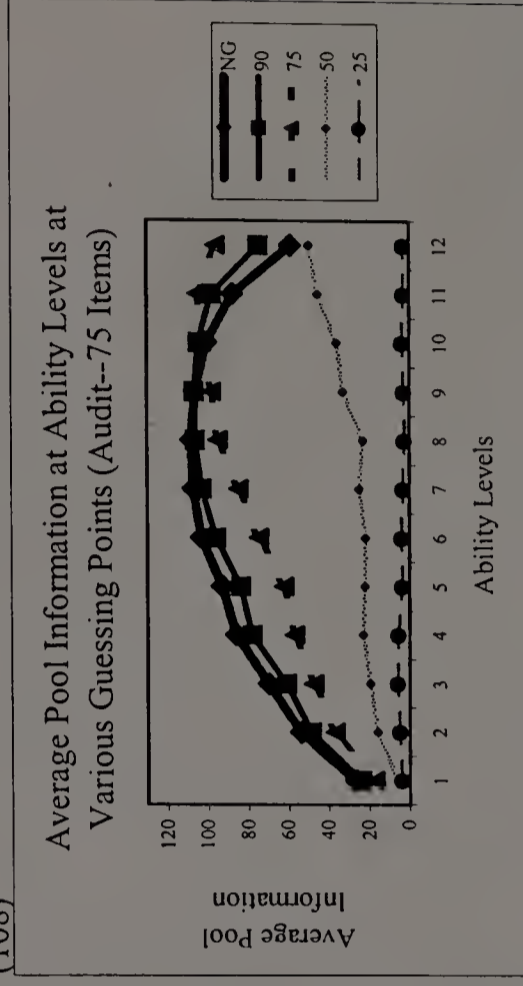
(106)



(107)

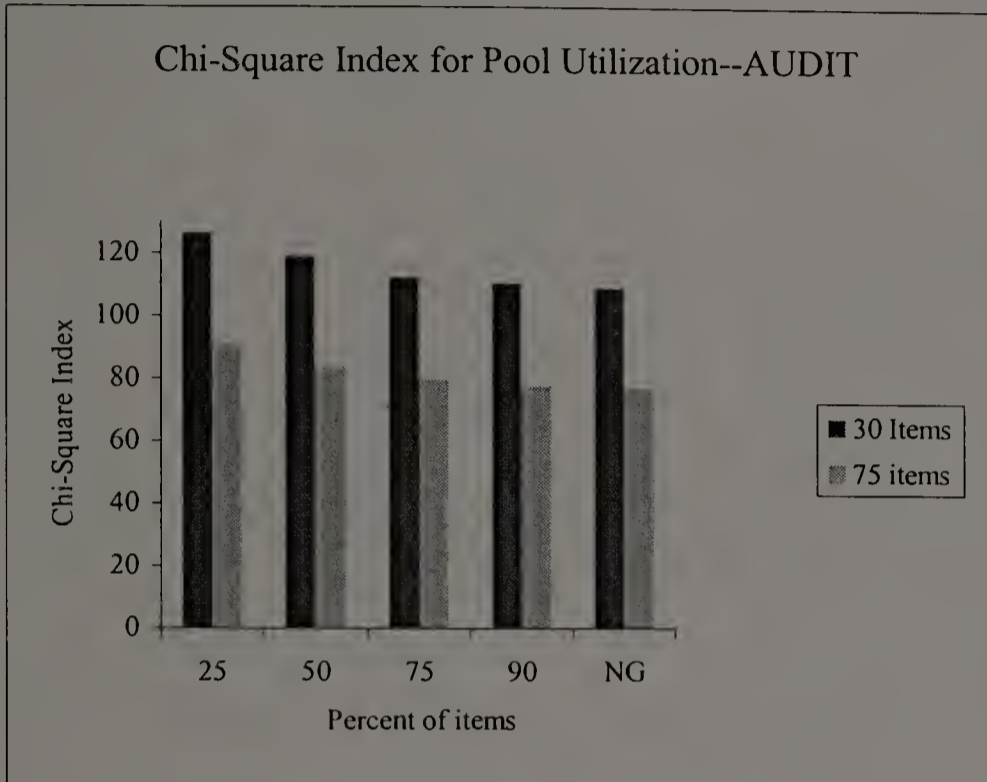


(108)



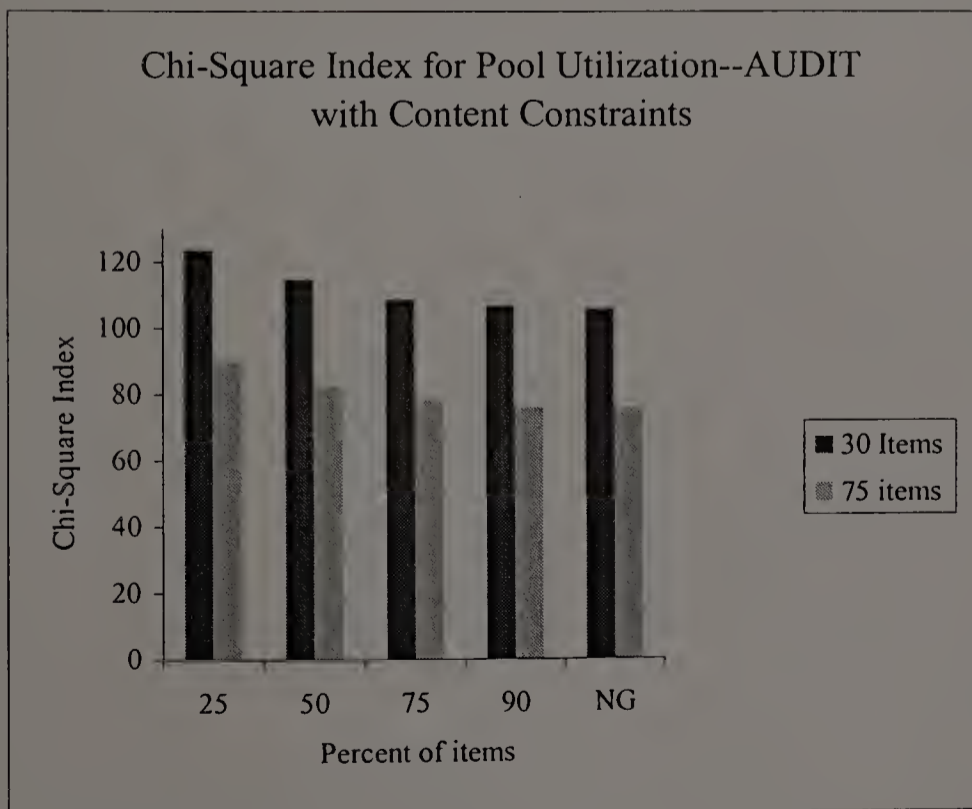
Pool Utilization Index at each Ability Level for 5 Guessing Scenarios  
 (Performance Testing with AICPA Parameters for 30 and 75 items on AUDIT  
 without/with Content Constraints)

(109)



	30 items	75 items
25	126.48	90.74
50	118.98	83.67
75	112.29	79.55
90	110.51	77.59
NG	108.90	76.67

(110)

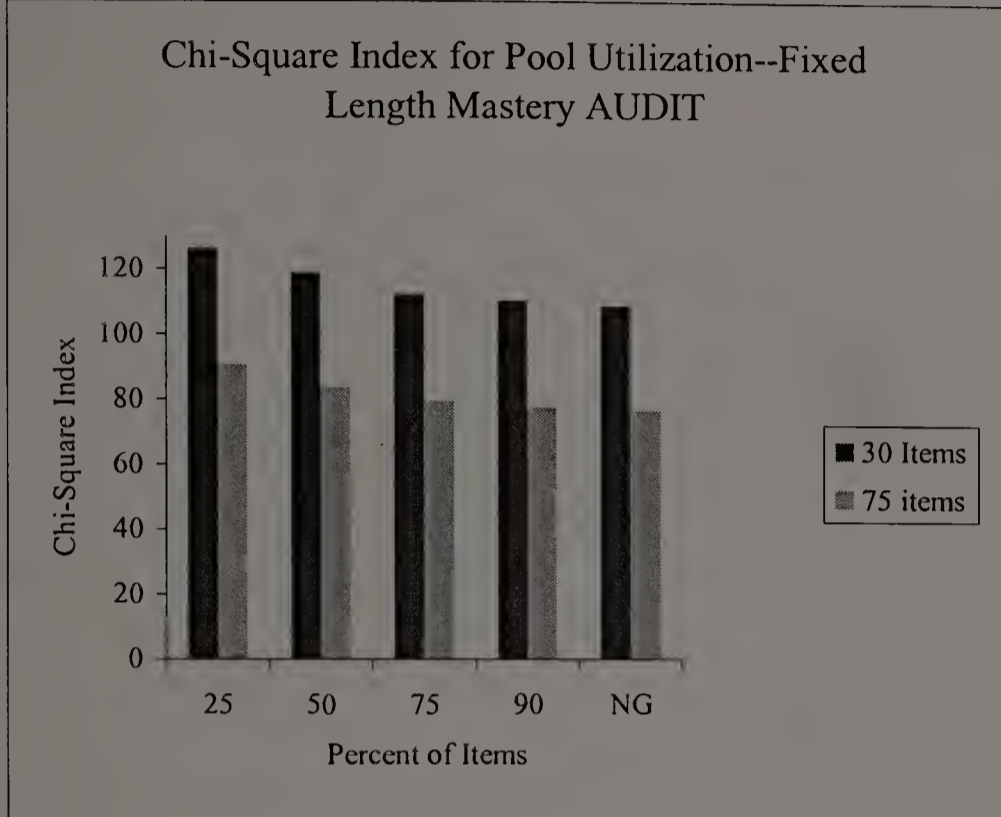


	30 items	75 items
25	123.71	89.99
50	114.81	82.55
75	108.70	78.26
90	106.77	75.92
NG	105.92	75.79



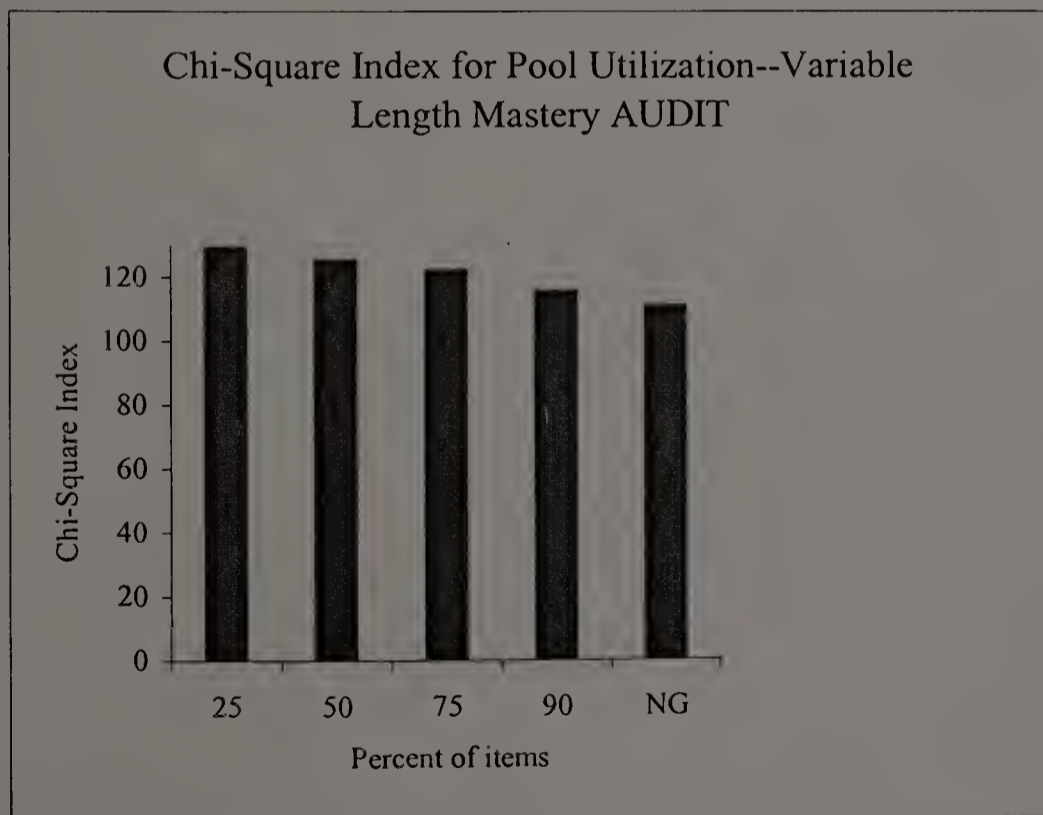
Pool Utilization Index at each Ability Level for 5 Guessing Scenarios  
 (Mastery Testing with AICPA Parameters for Fixed/Variable Length AUDIT)

(111)



	30 items	75 items
25	126.48	90.74
50	118.98	83.67
75	112.29	79.55
90	110.51	77.59
NG	108.90	76.67

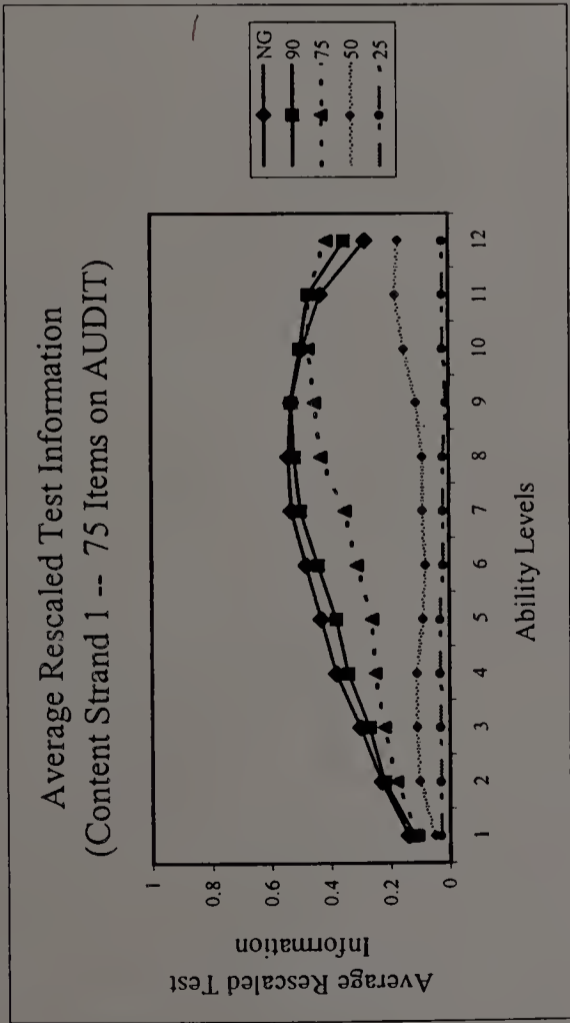
(112)



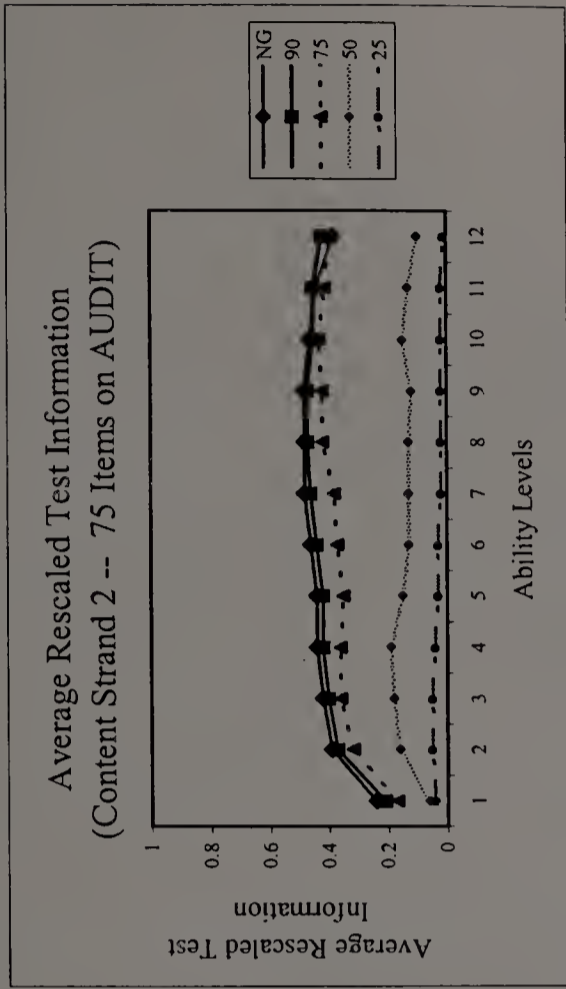
Variable Length	
25	129.98
50	125.85
75	122.66
90	116.10
NG	111.52

Average Rescaled Test Information in each Content Area at each Ability Level  
 (Performance Testing with AICPA Parameters for 75 Items on AUDIT)

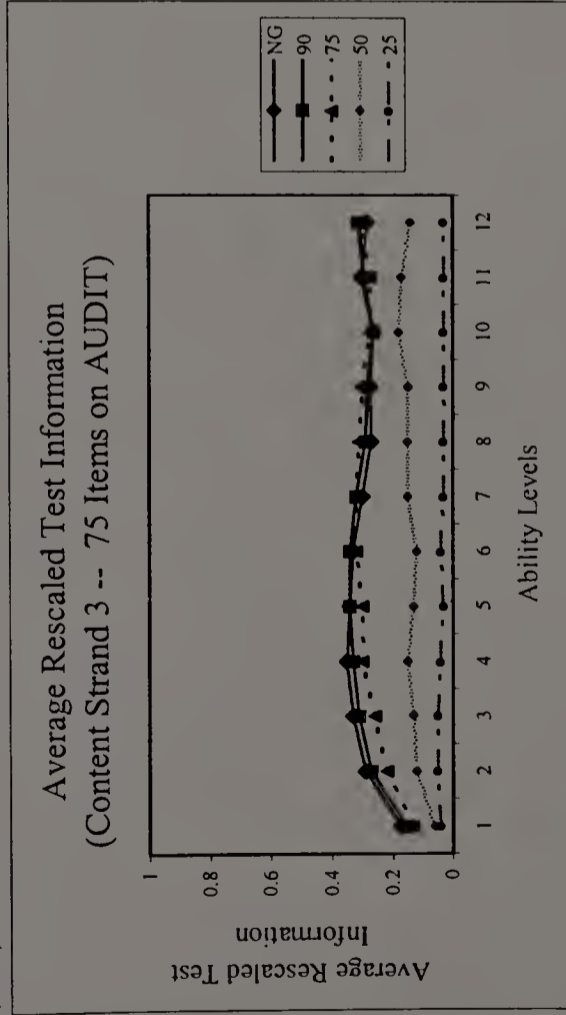
(113)



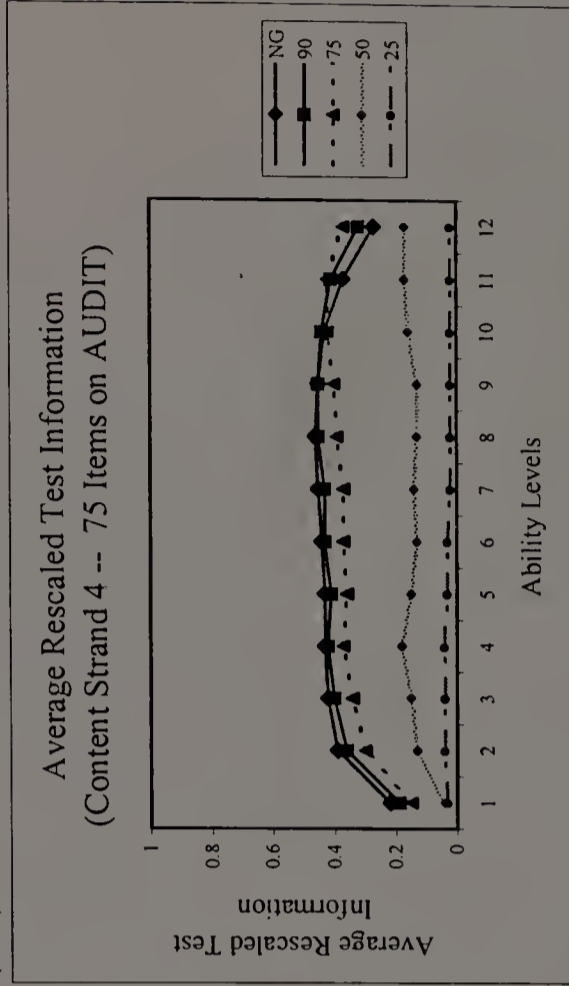
(114)



(115)

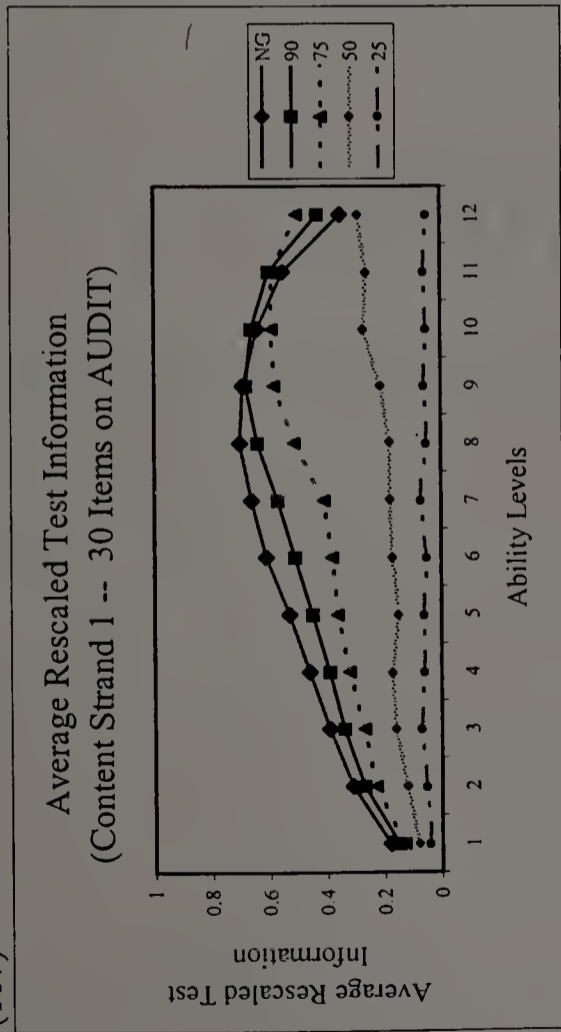


(116)

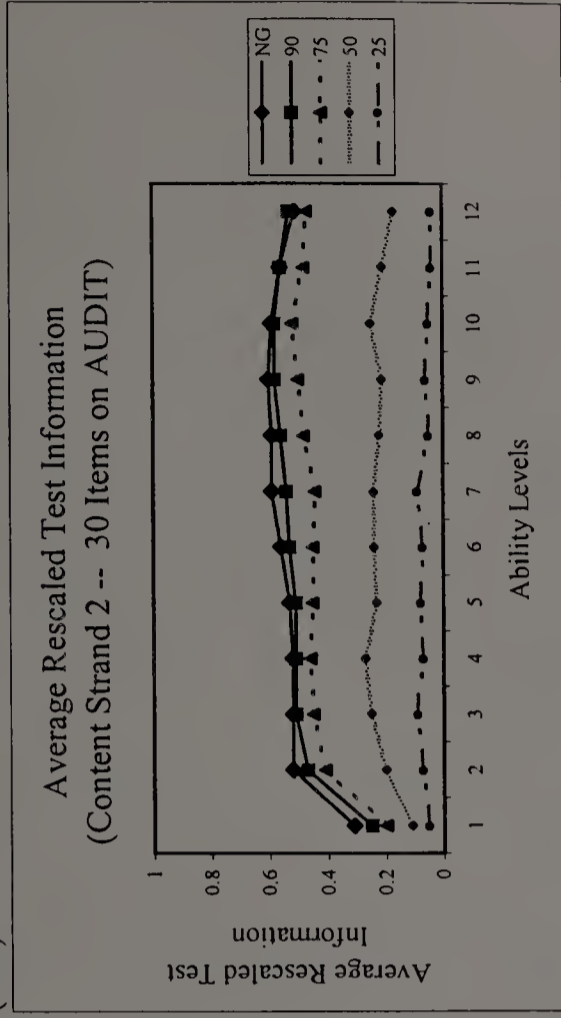


Average Rescaled Test Information in each Content Area at each Ability Level  
 (Performance Testing with AICPA Parameters for 30 Items on AUDIT)

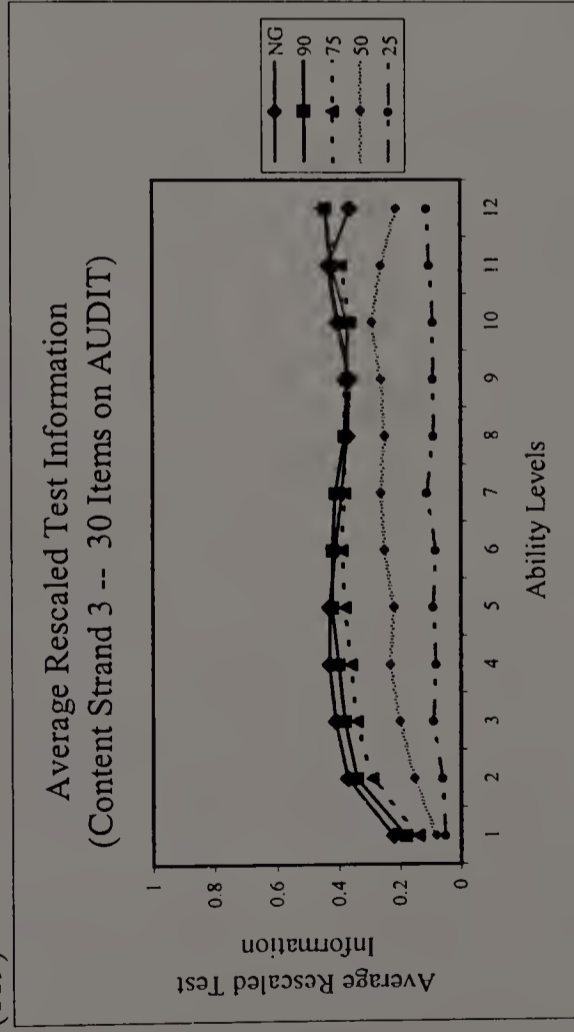
(117)



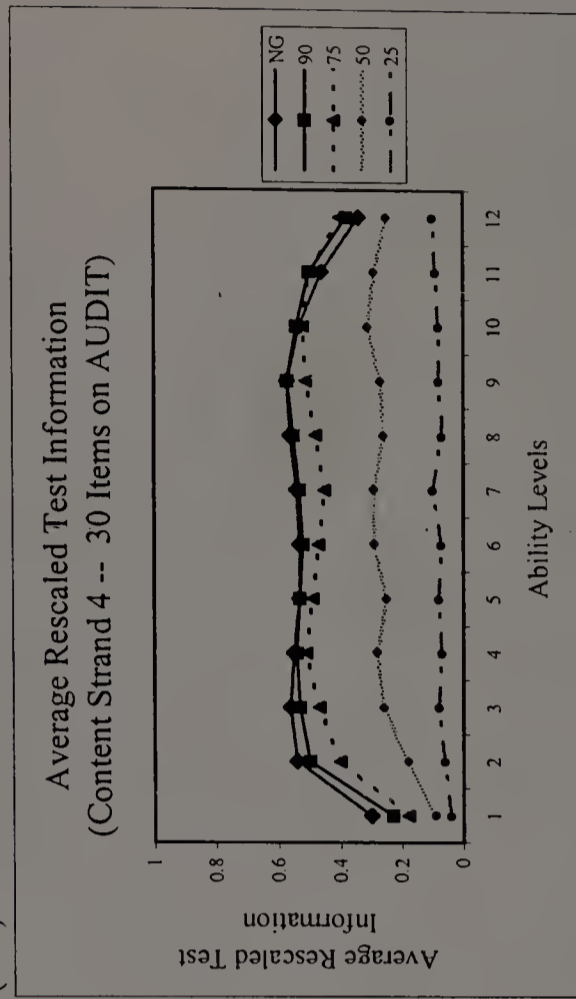
(118)



(119)

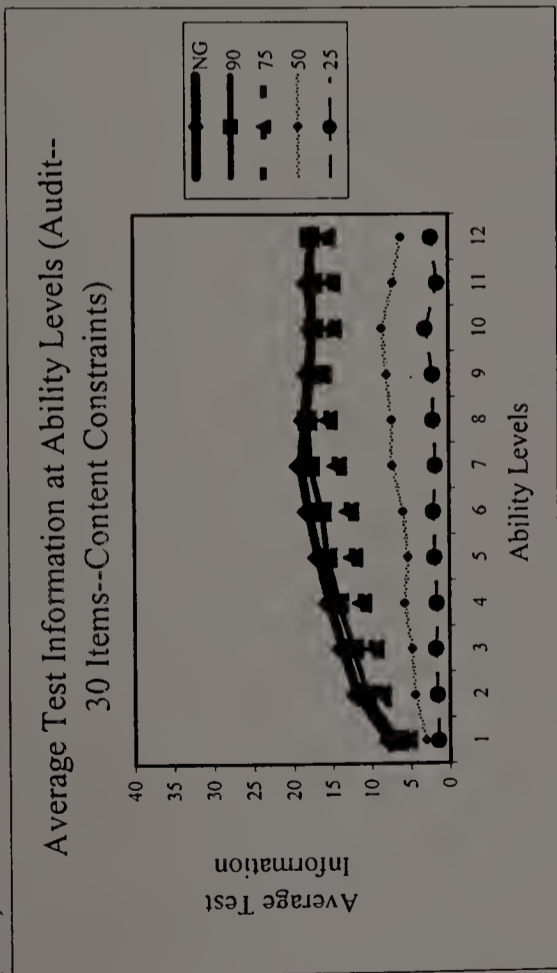


(120)

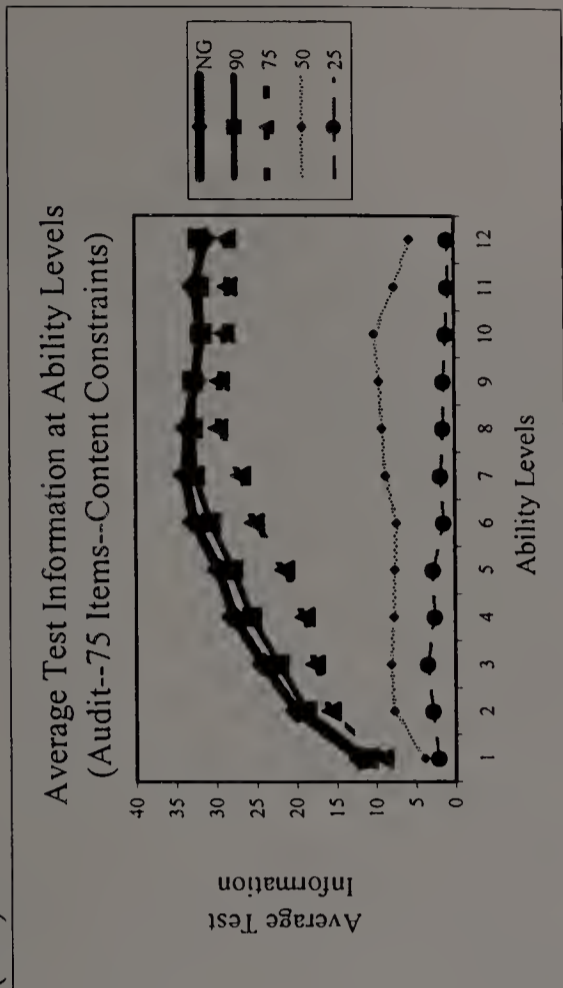


Average Test and Pool Information at each Ability Level for 5 Guessing Scenarios  
 (Performance Testing with AICPA Parameters for 30 and 75 Items on AUDIT with Content Constraints)

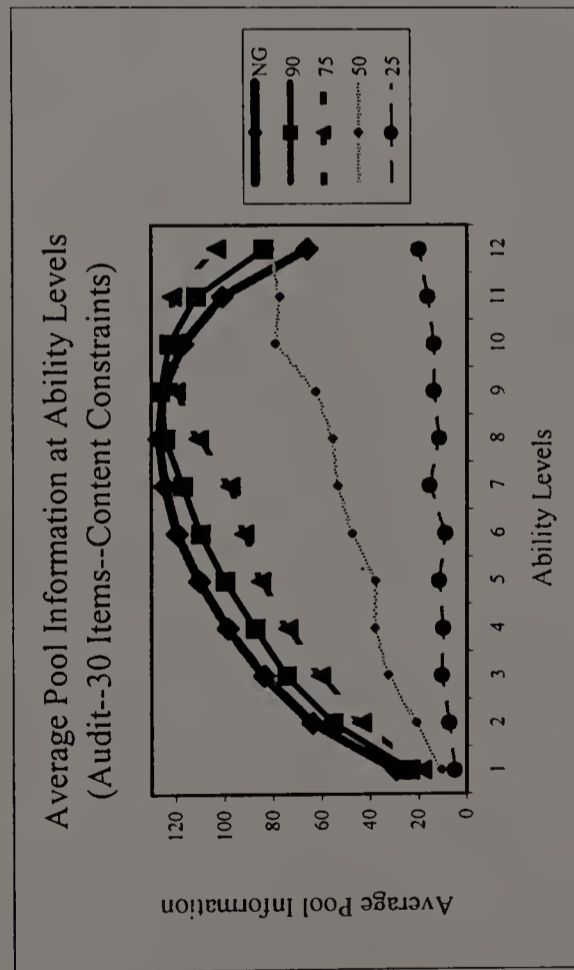
(121)



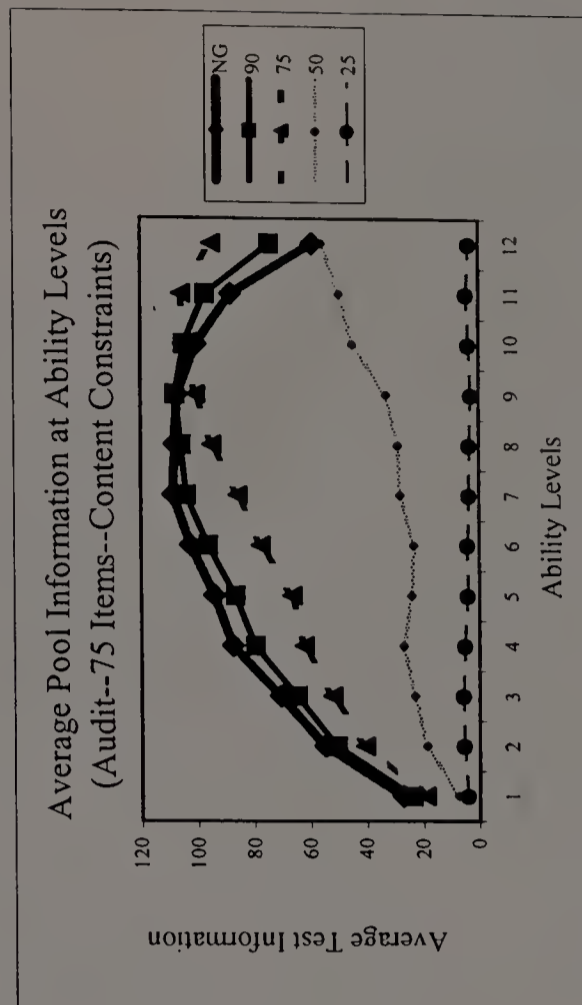
(122)



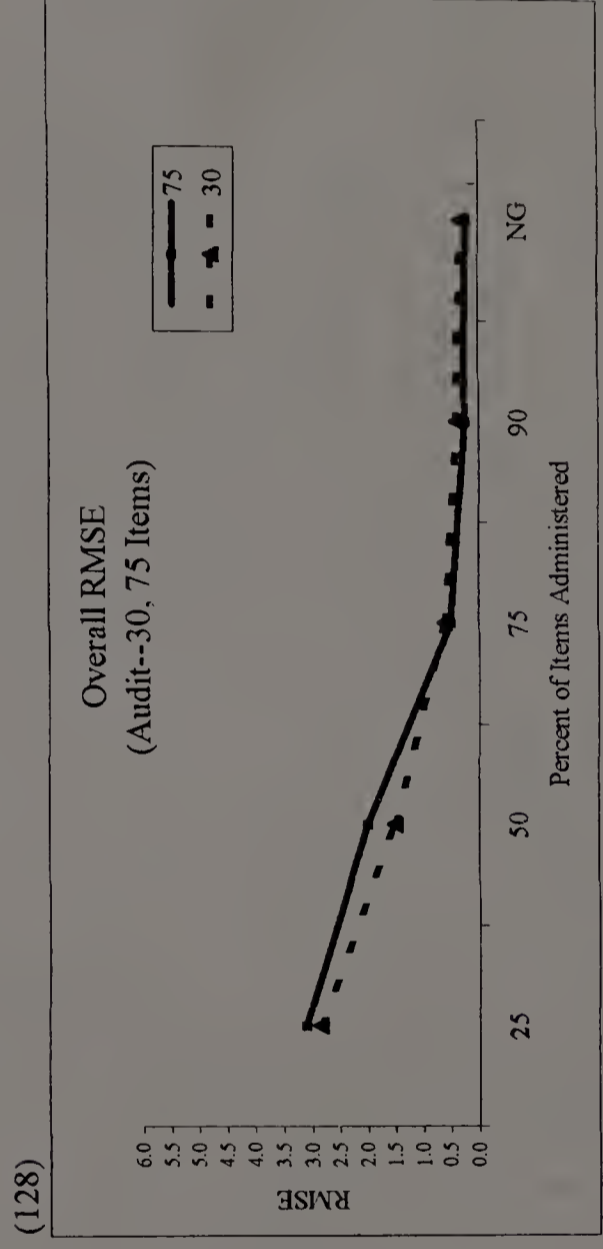
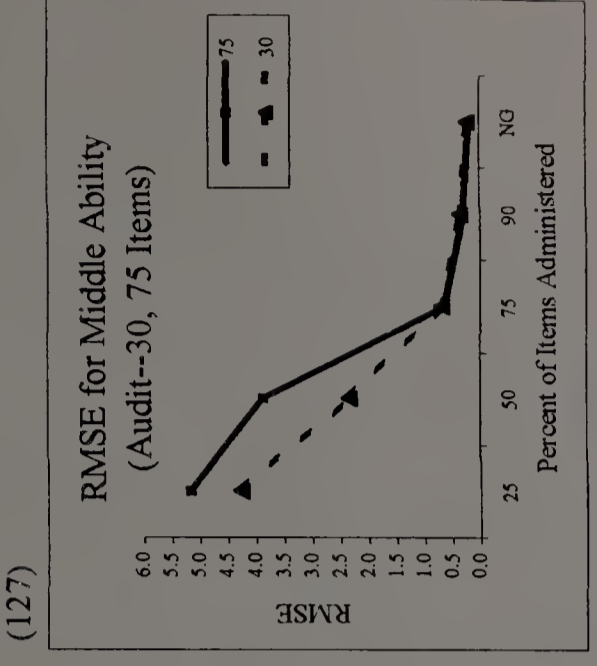
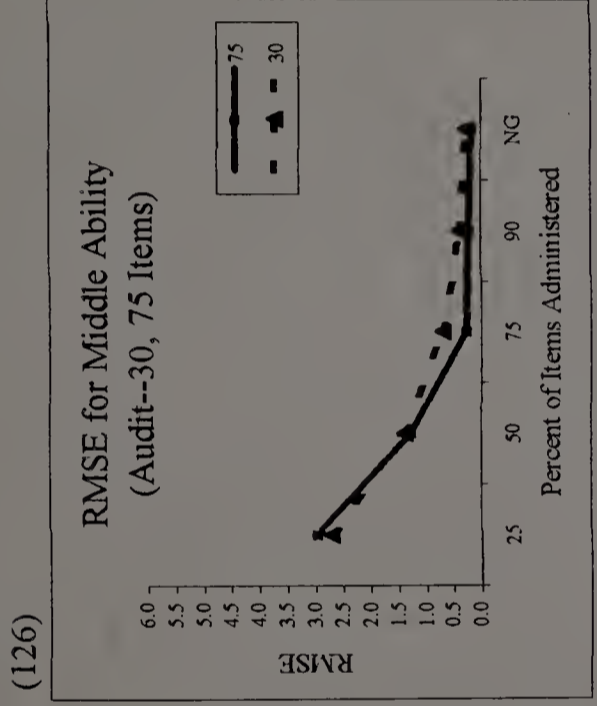
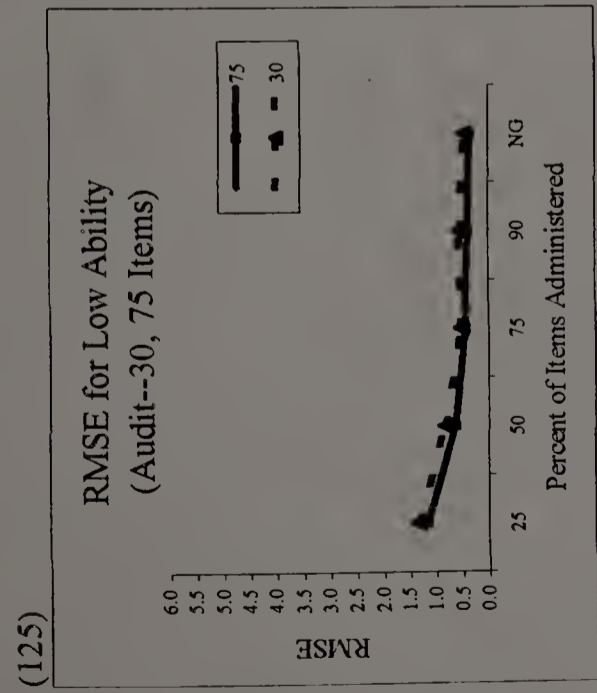
(123)



(124)

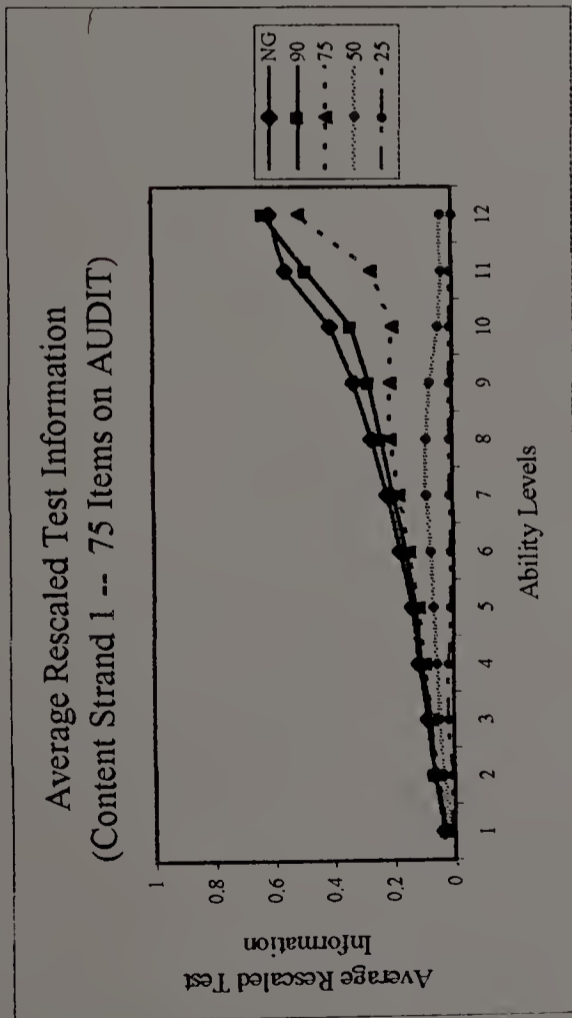


RMSE of Estimates around True Ability for 5 Guessing Scenarios  
 (Performance Testing with AICPA Parameters for 30 and 75 Items on AUDIT with Content Constraints  
 b-Parameter Increased by a Constant for Content Strand 1)

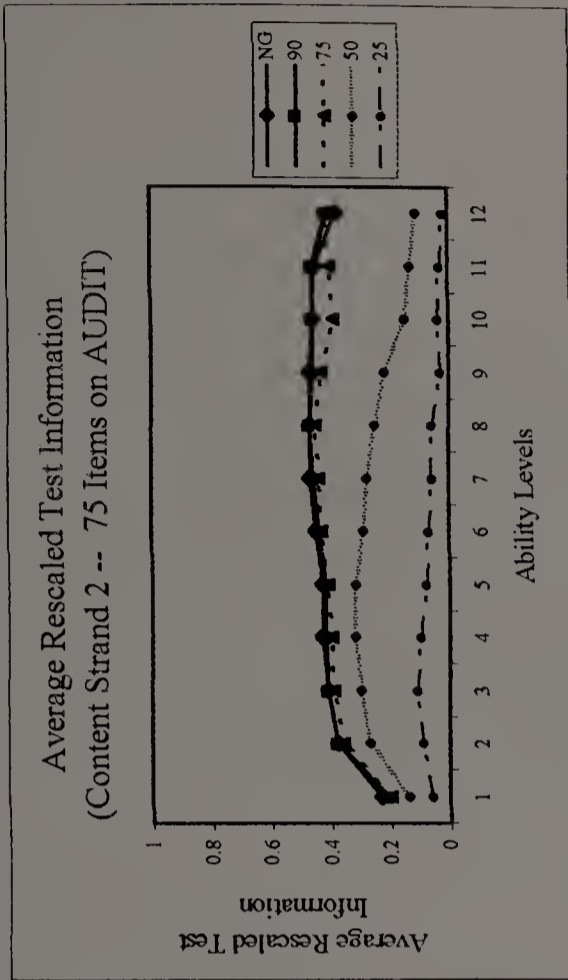


Average Test and Pool Information at each Ability Level for 5 Guessing Scenarios  
 (Performance Testing with AICPA Parameters for 75 Items on AUDIT with Content Constraints)  
 b-Parameter Increased by a Constant for Content Strand 1)

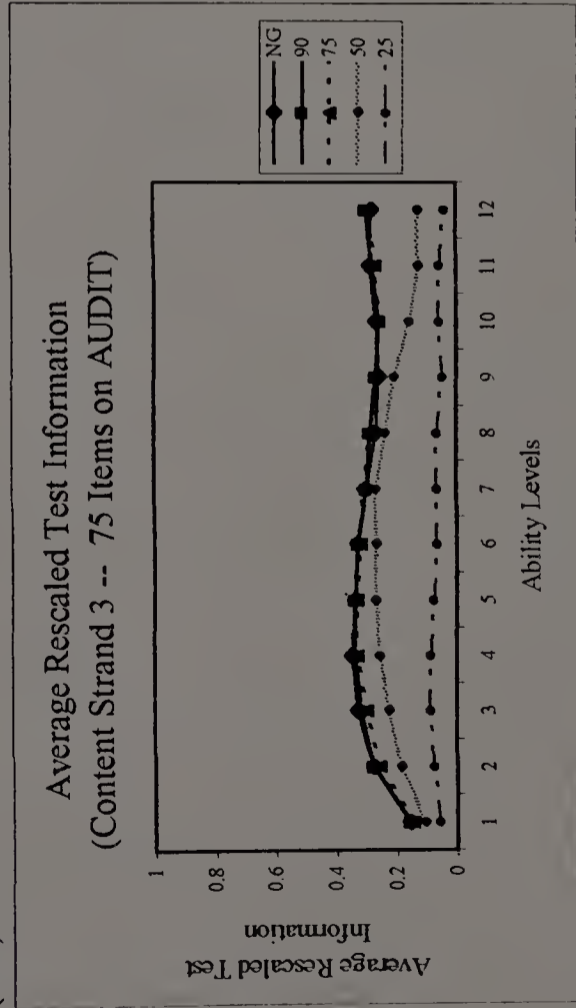
(129)



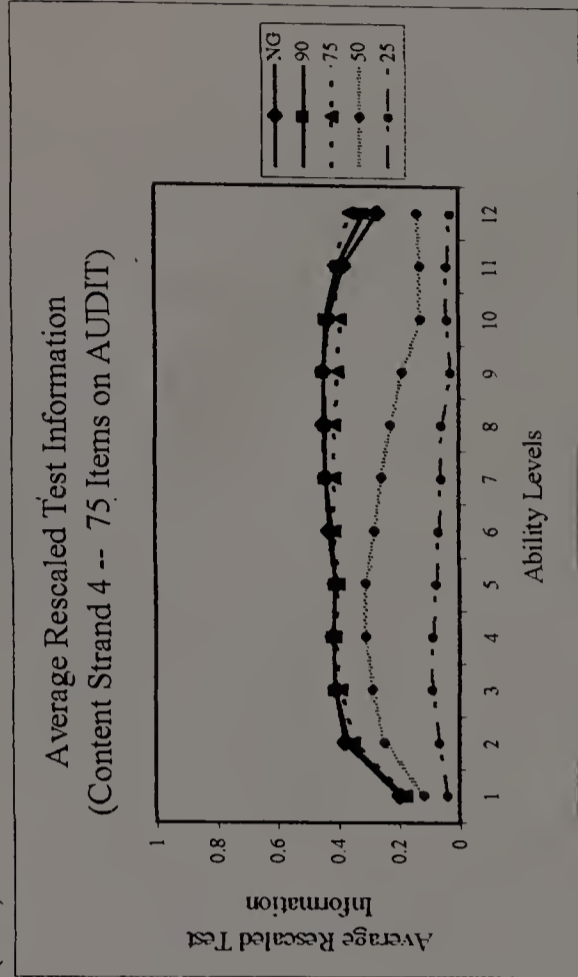
(130)



(131)

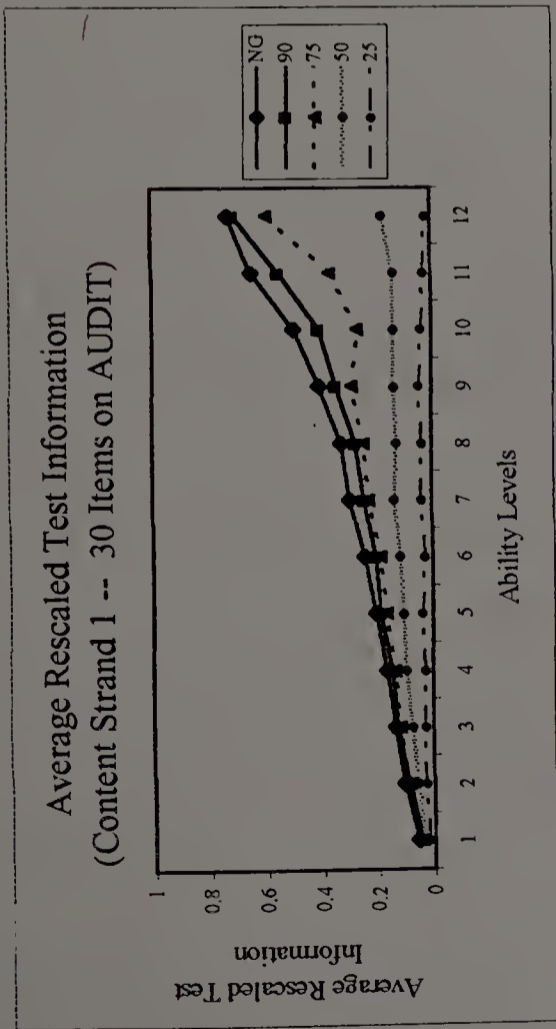


(132)

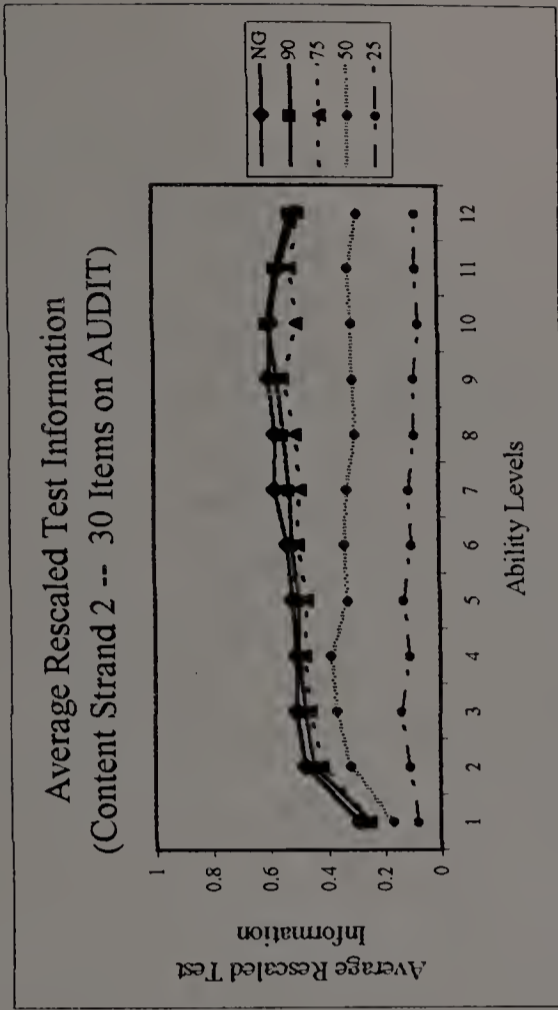


Average Rescaled Test Information at each Ability Level for 5 Guessing Scenarios  
 (Performance Testing with AICPA Parameters for 30 Items on AUDIT with Content Constraints)  
 b-Parameter Increased by a Constant for Content Strand 1)

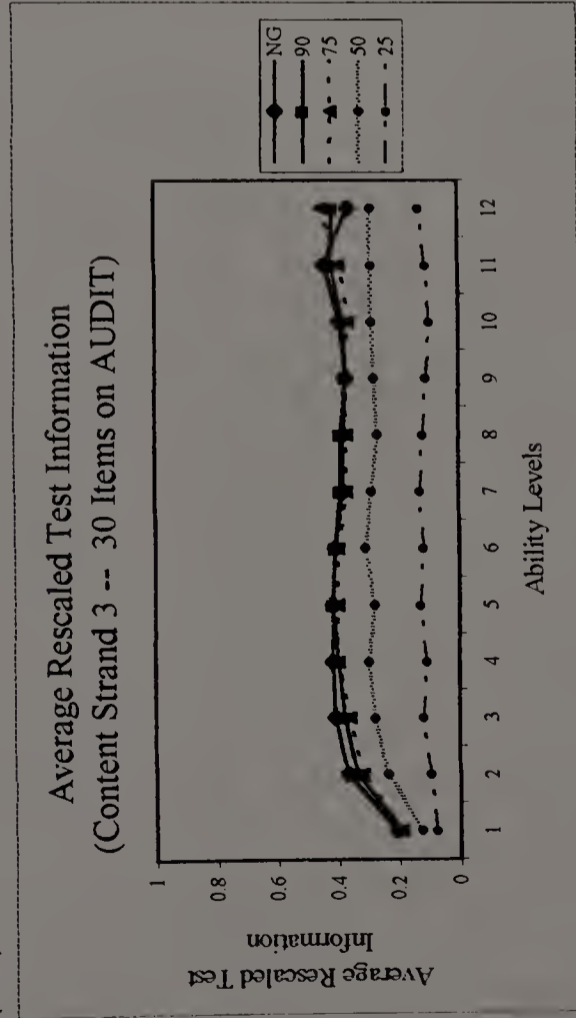
(133)



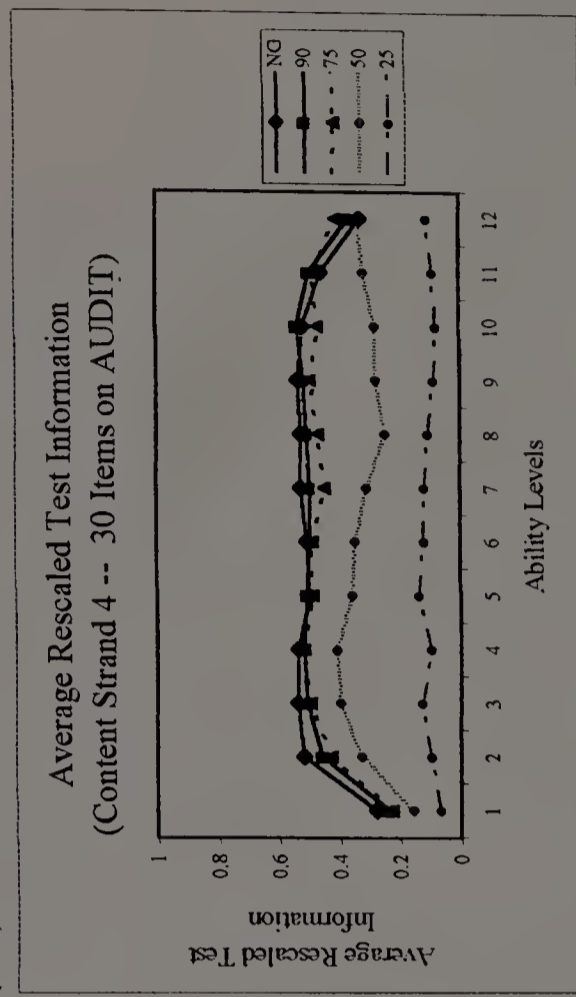
(134)



(135)

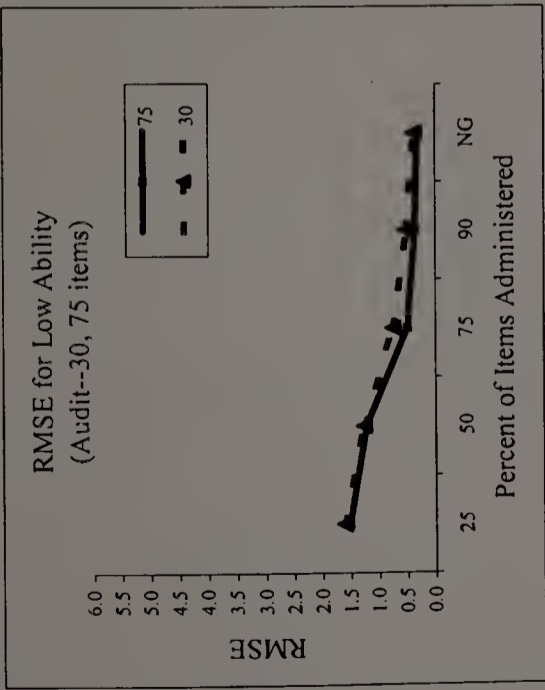


(136)

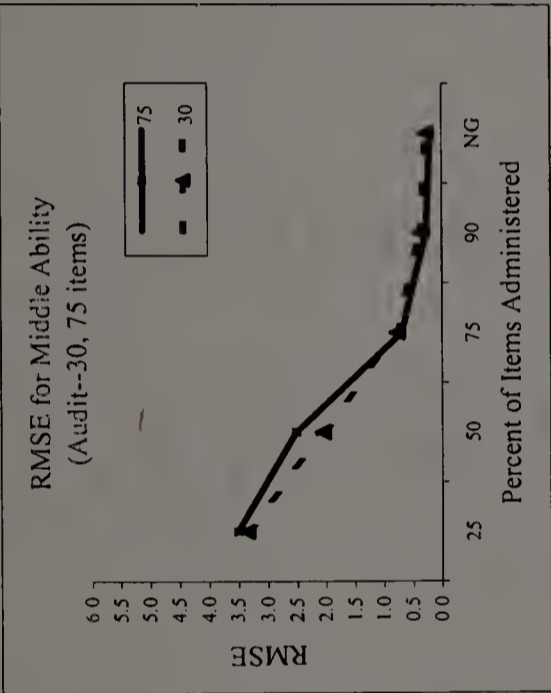


RMSE of Estimates around True Ability for 5 Guessing Scenarios  
(Fixed Length Mastery Testing with AICPA Parameters for 30 and 75 Items on AUDIT with Content Constraints)

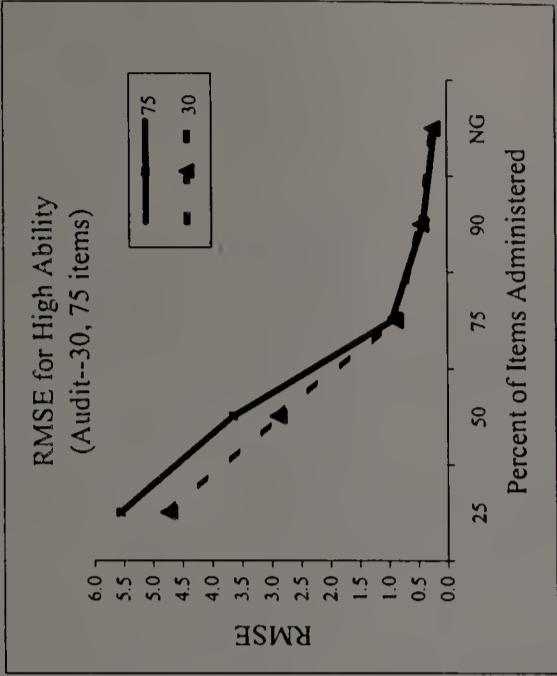
(137)



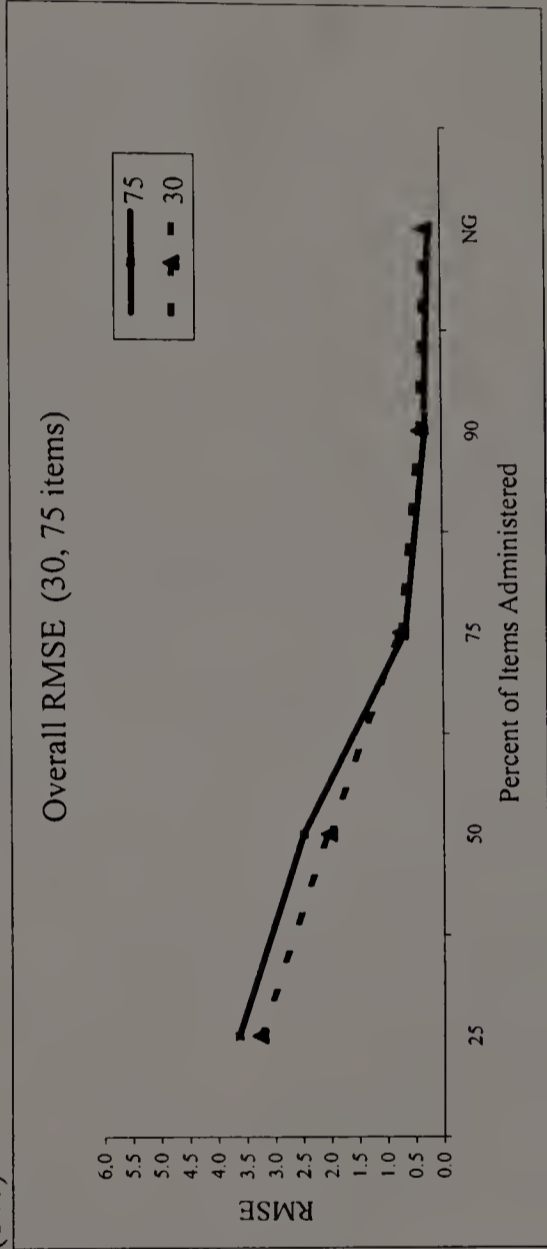
(138)



(139)



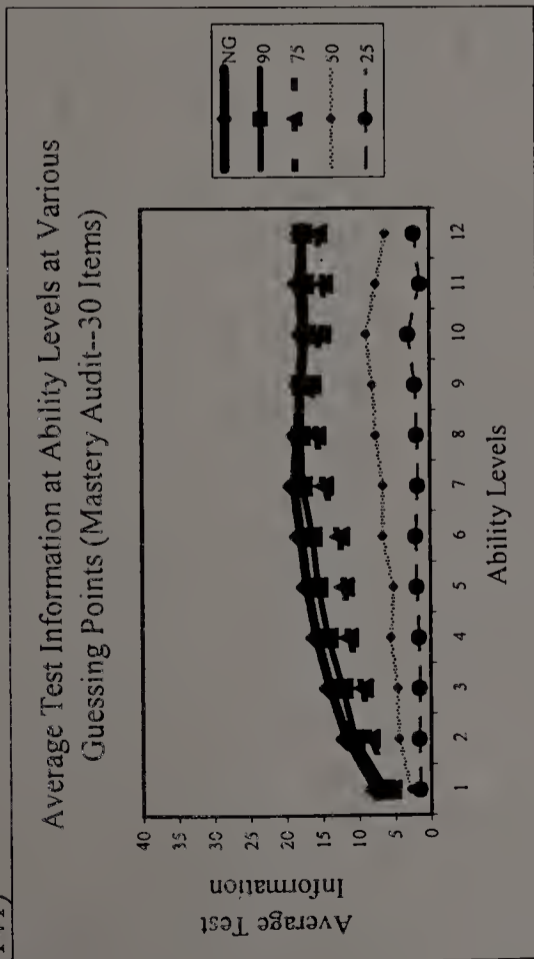
(140)



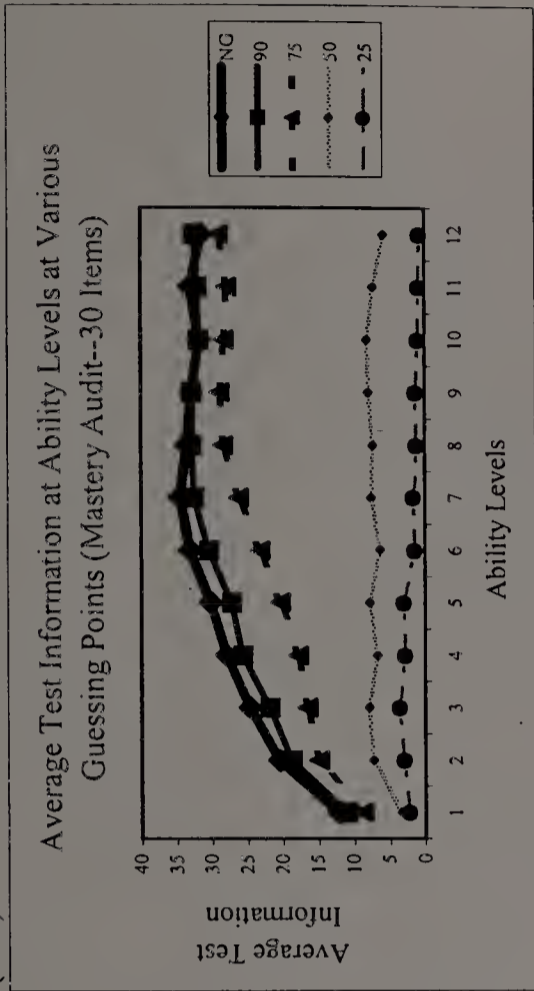


Average Test and Pool Information at each Ability Level for 5 Guessing Scenarios  
 (Fixed Length Mastery Testing with AICPA Parameters for 30 and 75 Items on AUDIT)

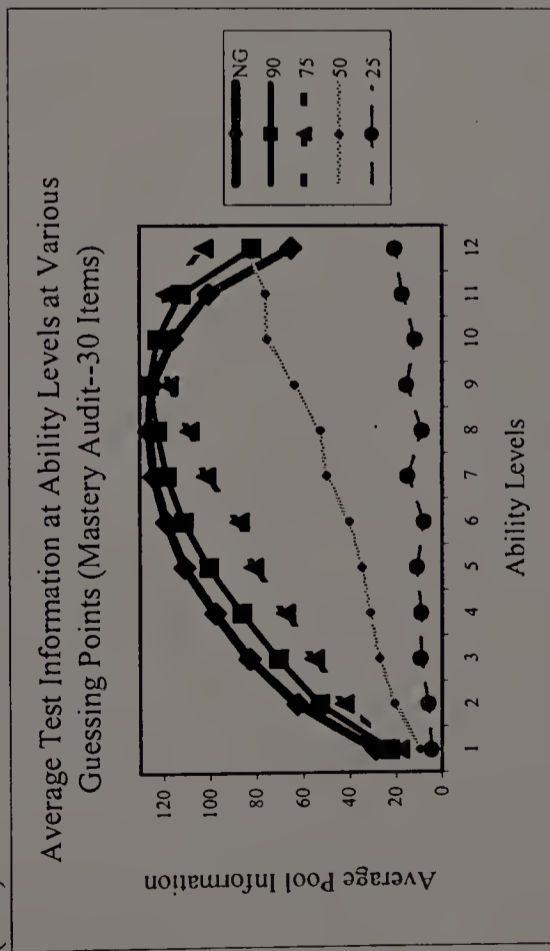
(141)



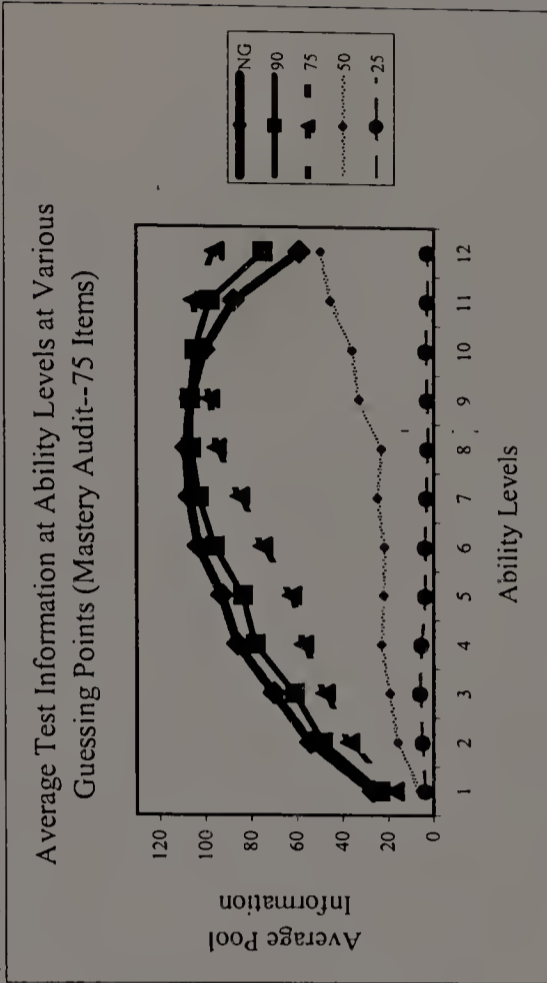
(142)



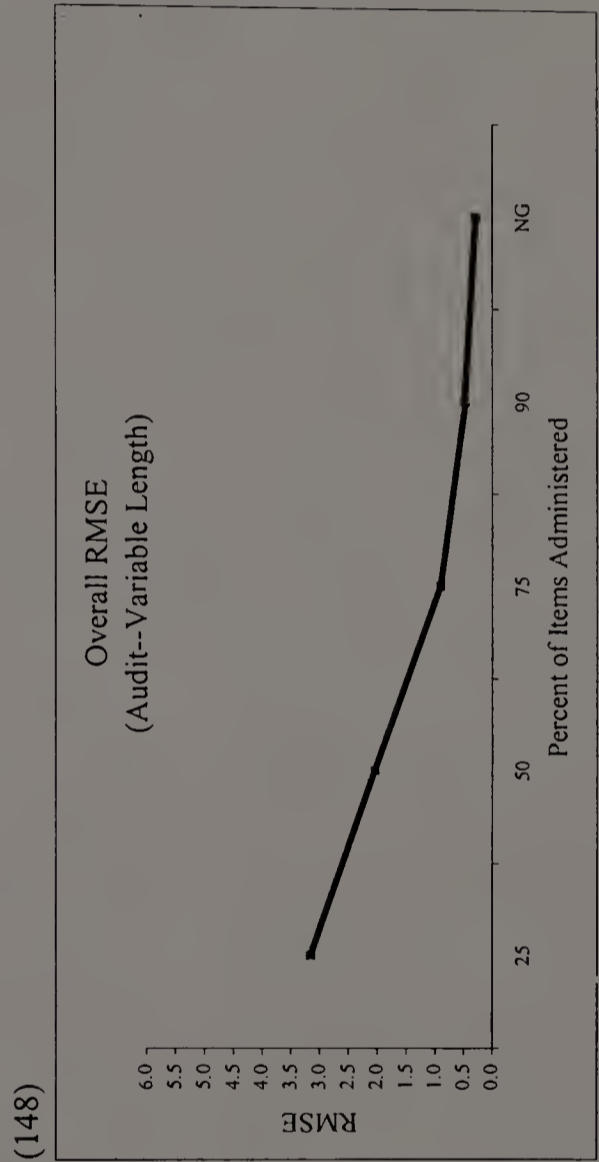
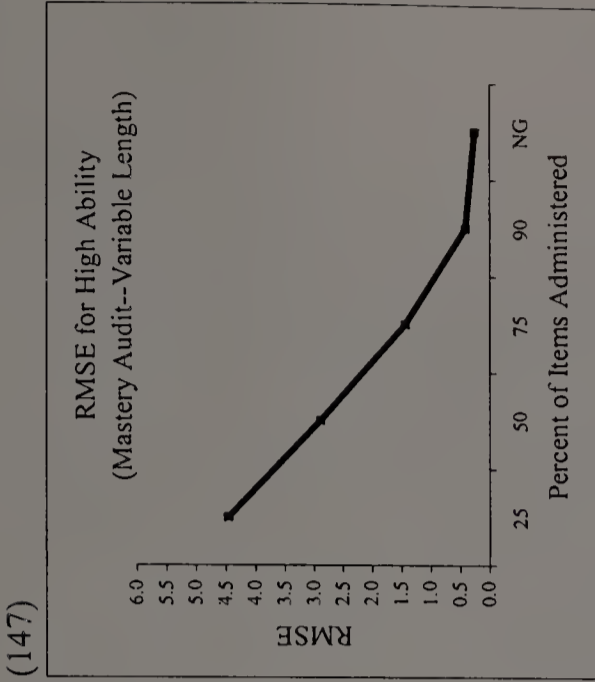
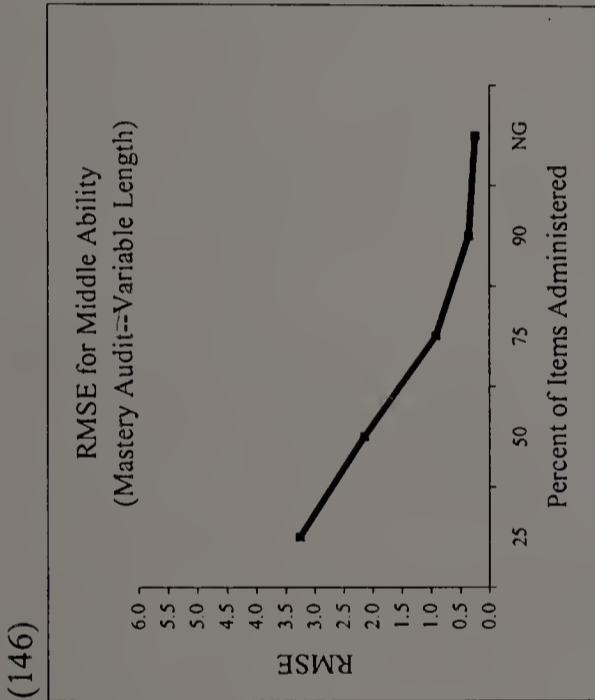
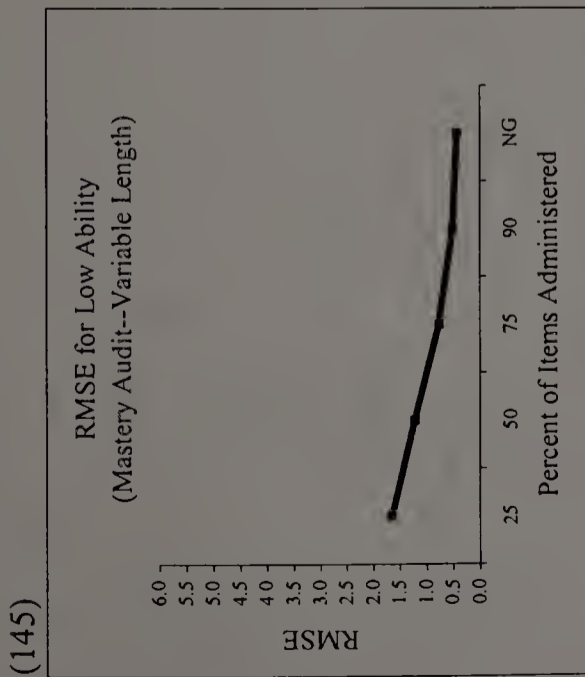
(143)



(144)

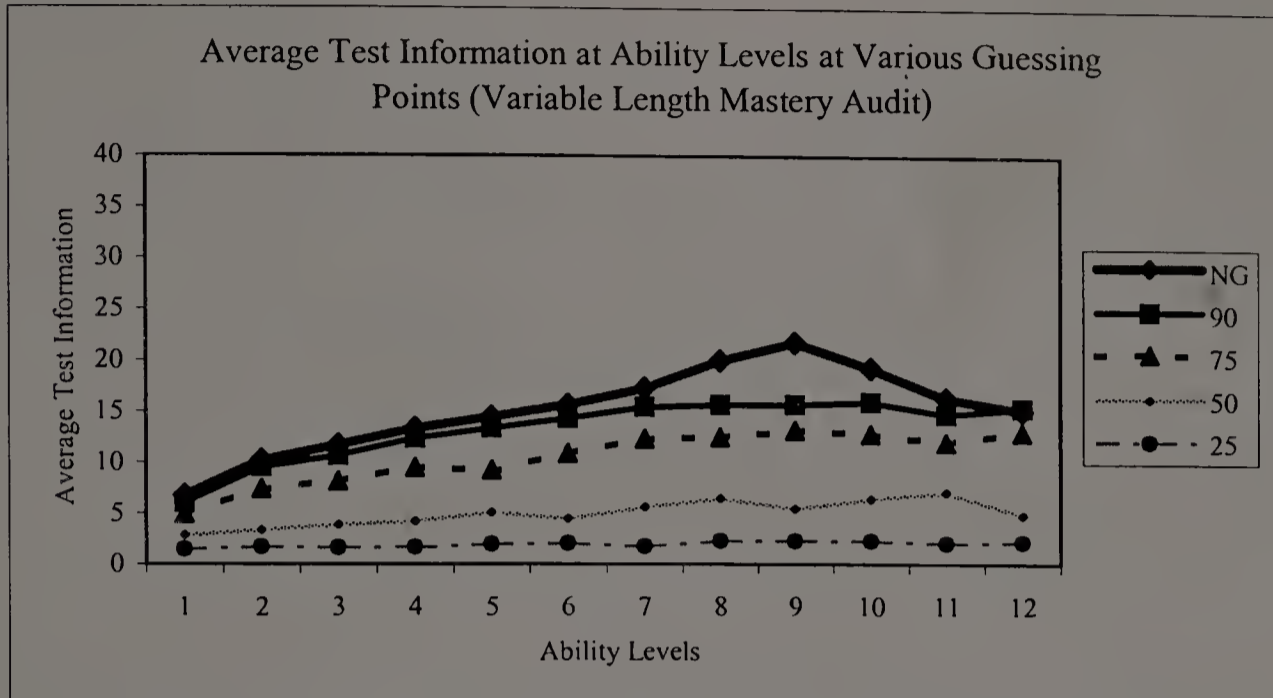


RMSE of Estimates around True Ability for 5 Guessing Scenarios  
 (Variable Length Mastery Testing with AICPA Parameters for AUDIT)

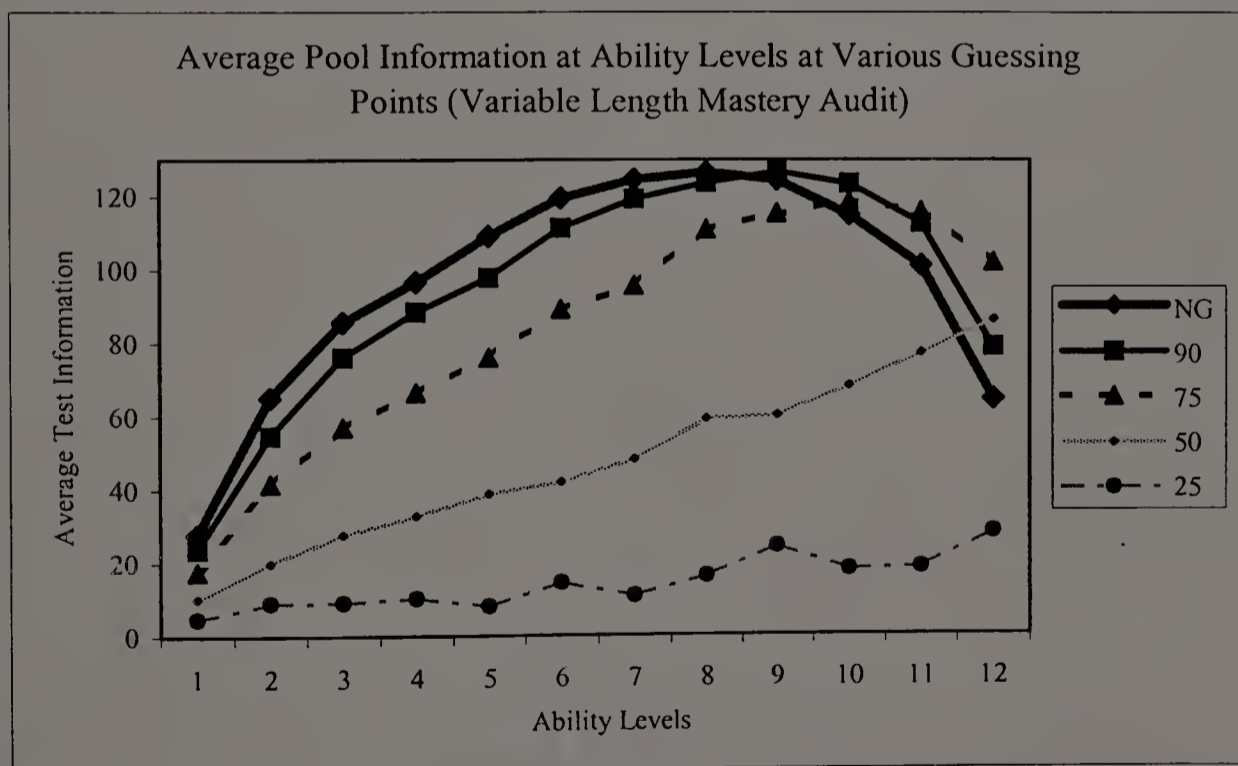


Average Test and Pool Information at each Ability Level for 5 Guessing Scenarios  
 (Variable Length Mastery Testing with AICPA Parameters -- AUDIT)

(149)



(150)

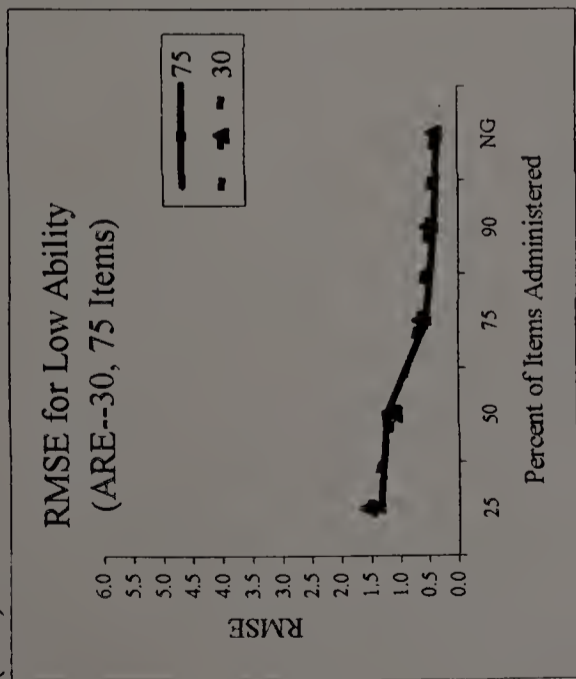


APPENDIX D

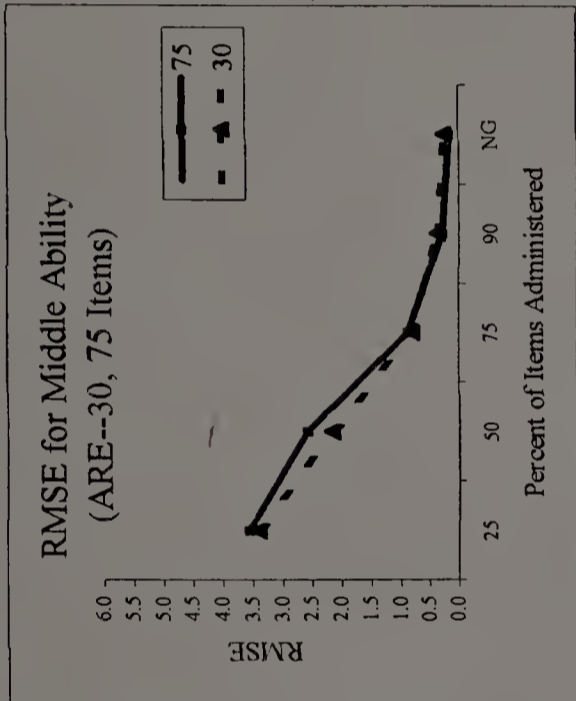
RESULTS USING AICPA ITEM PARAMETERS FOR ARE

RMSE of Estimates around True Ability for 5 Guessing Scenarios  
 (Performance Testing with AICPA Parameters for 30 and 75 Items on ARE)

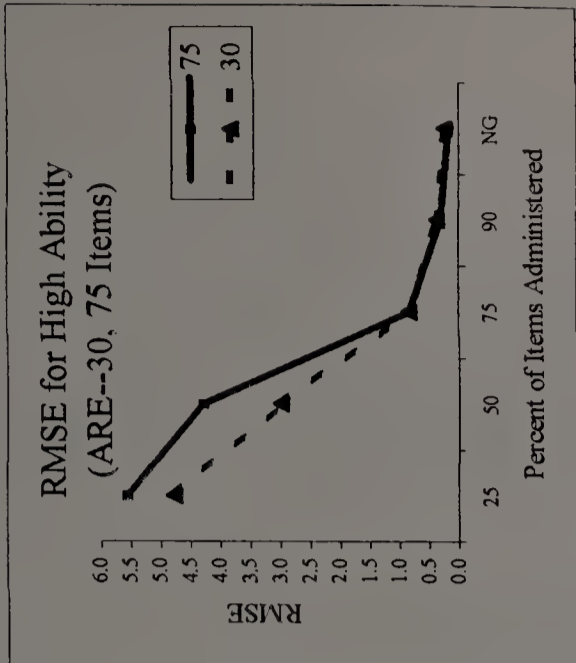
(151)



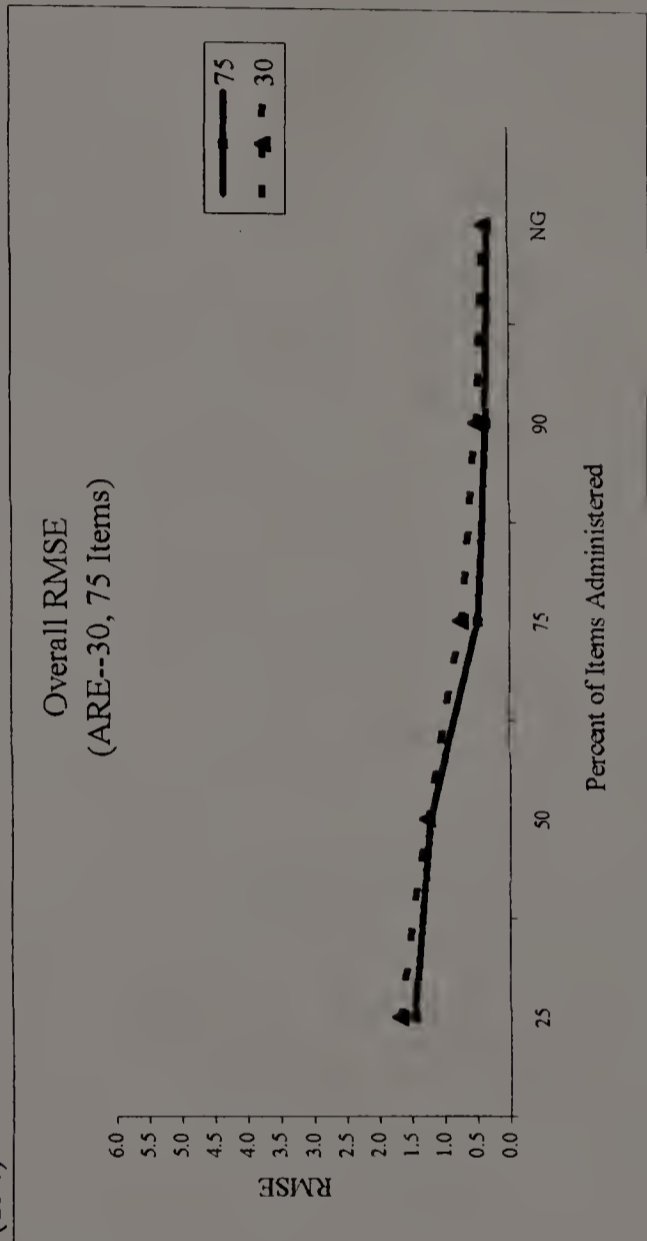
(152)



(153)

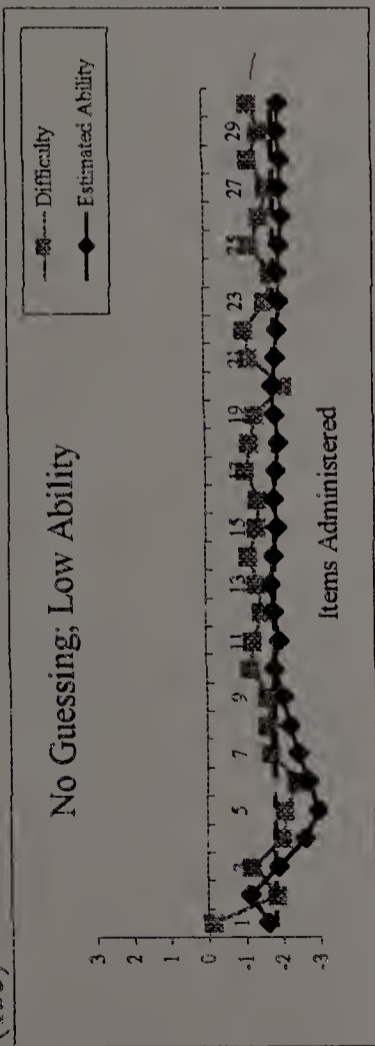


(154)

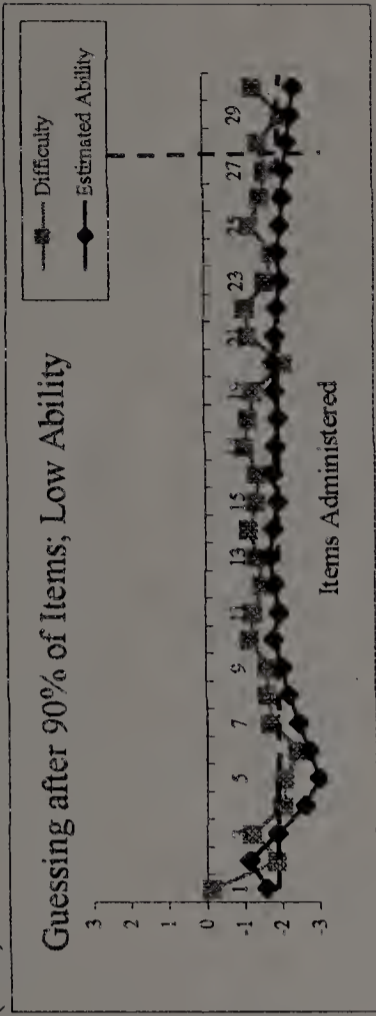


CAT Administration for a Low Ability Examinee who Guesses at a Certain Point in the Test (ARE--30 items)

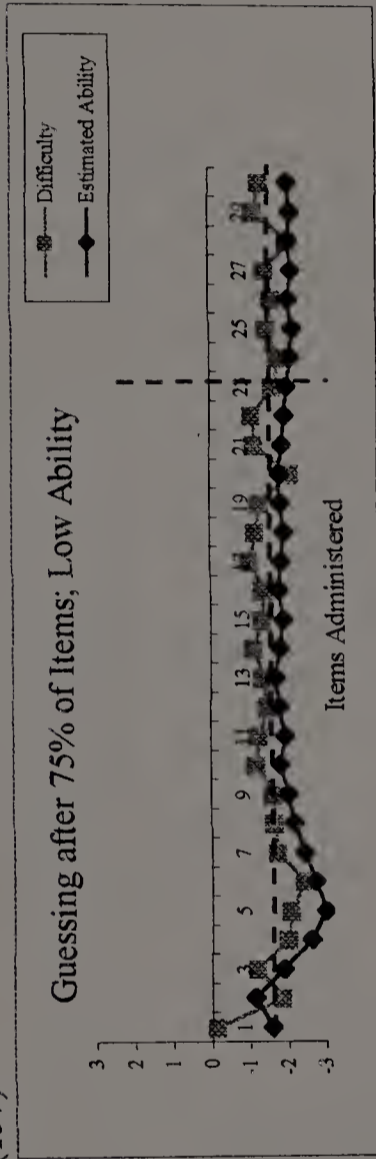
(155)



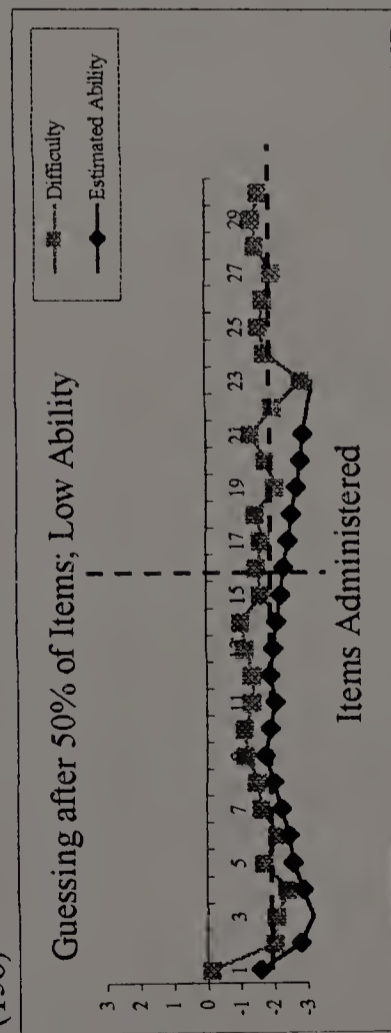
(156)



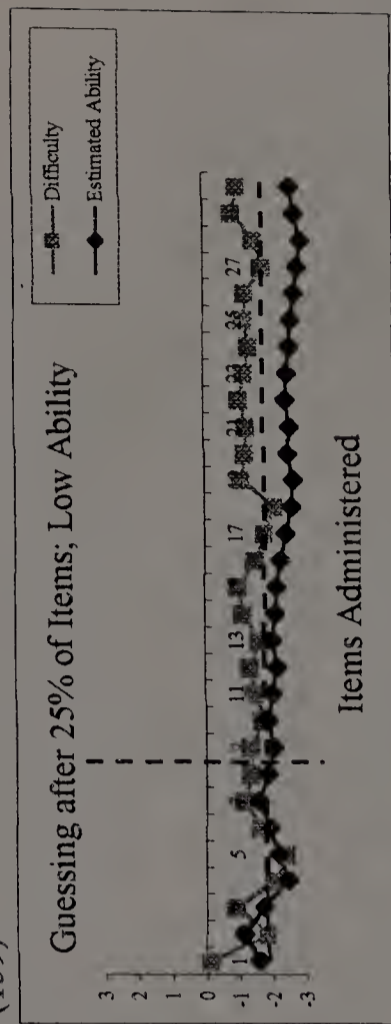
(157)



(158)

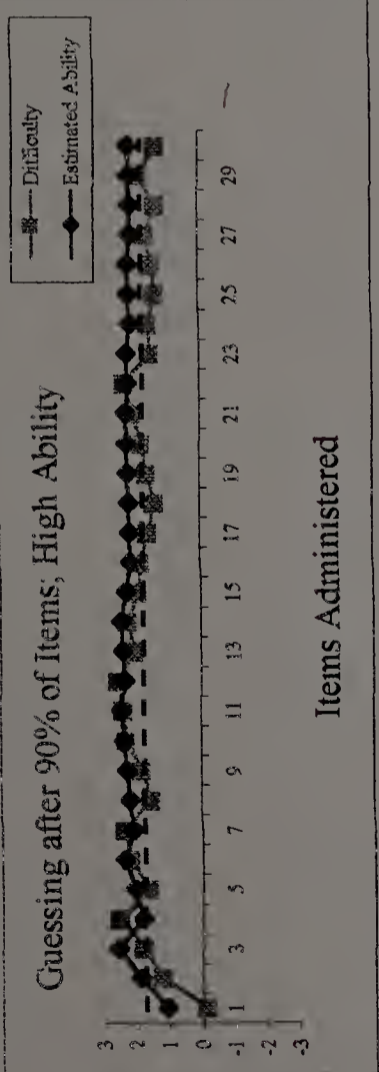


(159)

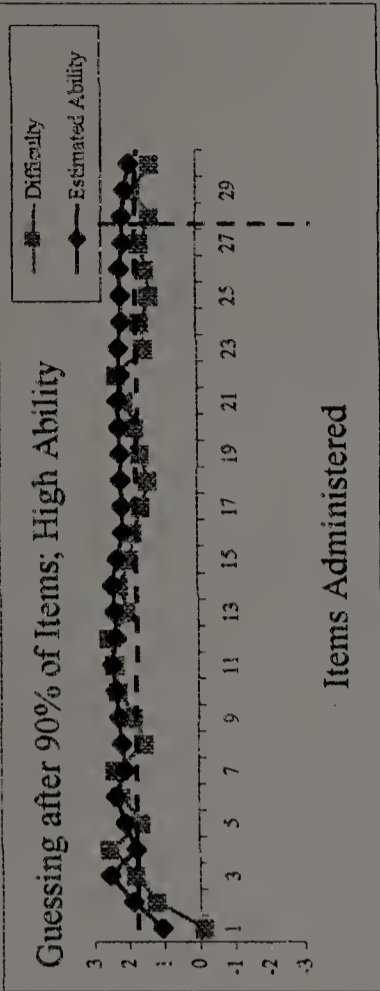


CAT Administration for a High Ability Examinee who Guesses at a Certain Point in the Test (ARE--30 items)

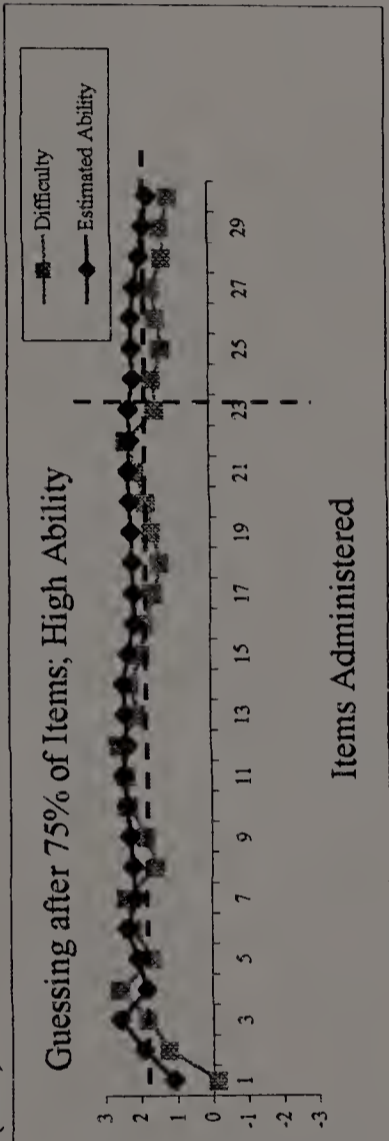
(165)



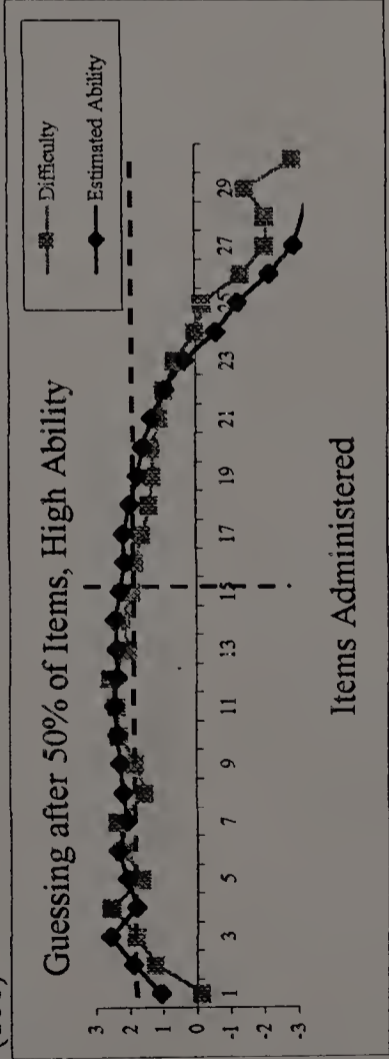
(166)



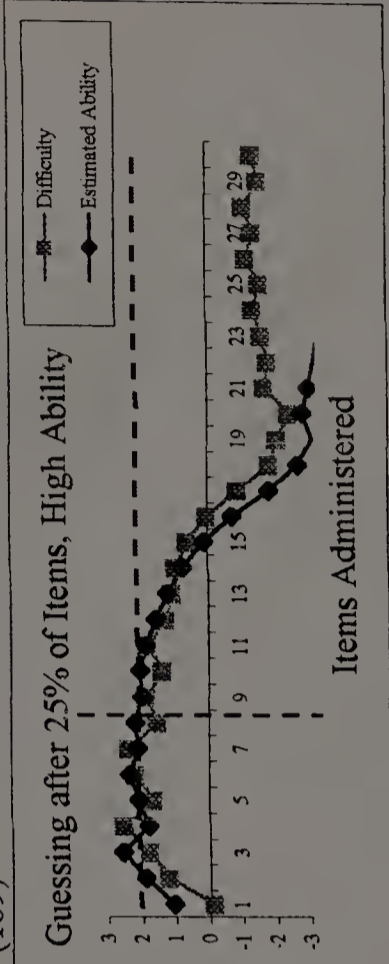
(167)



(168)

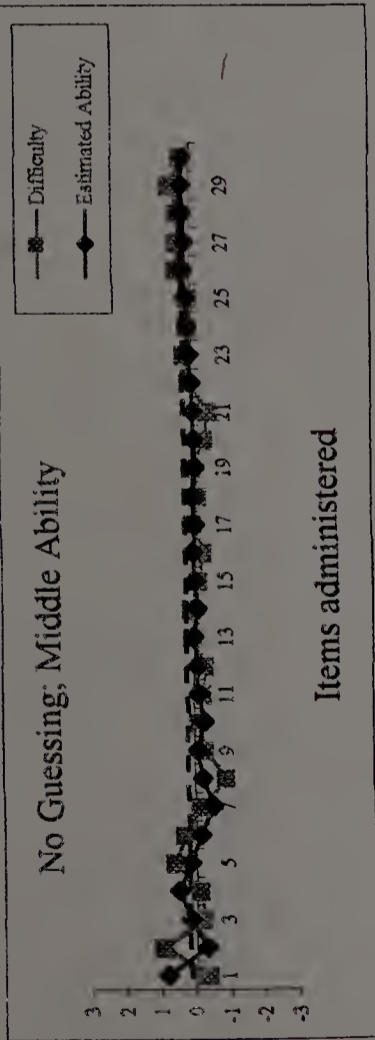


(169)

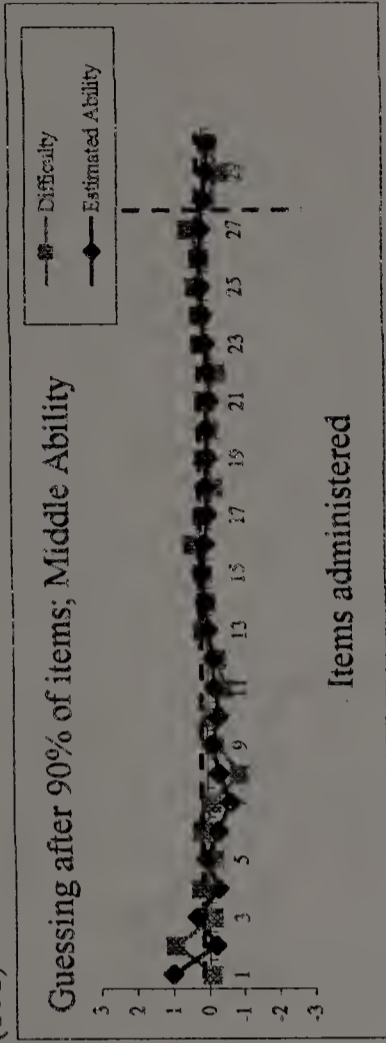


CAT Administration for a Middle Ability Examinee who Guesses at a Certain Point in the Test (ARE--30 items)

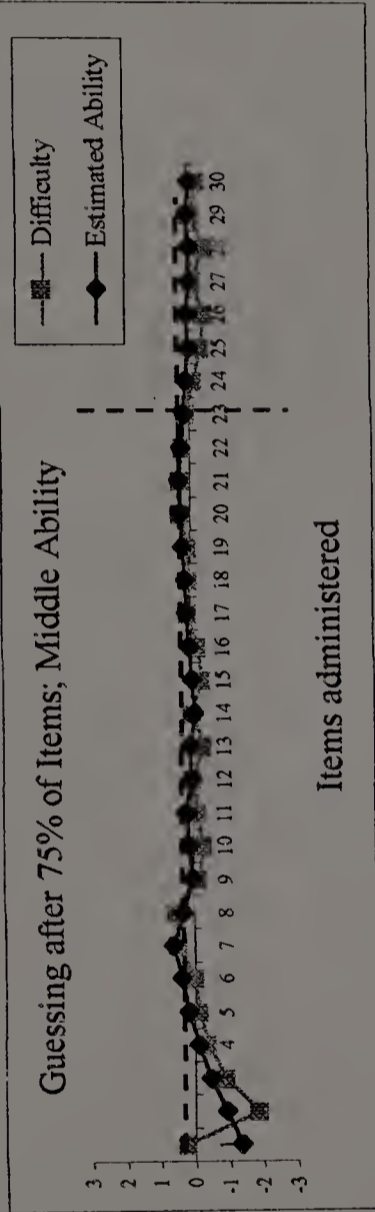
(160)



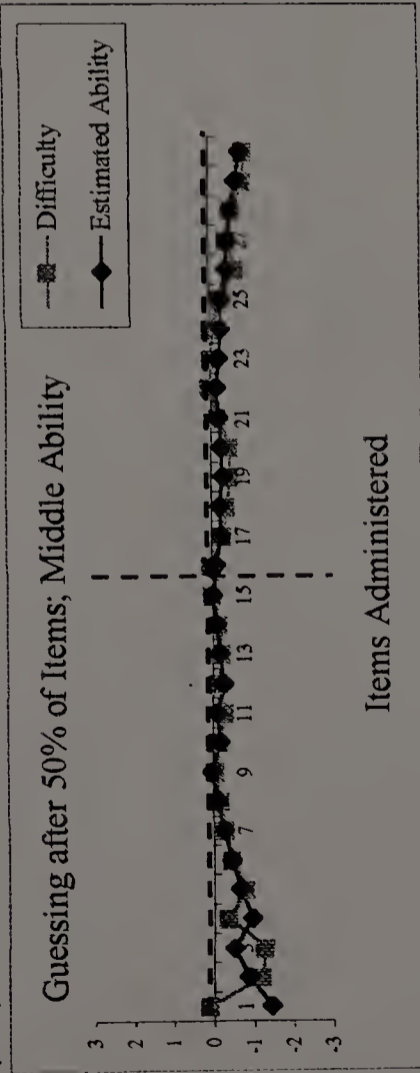
(161)



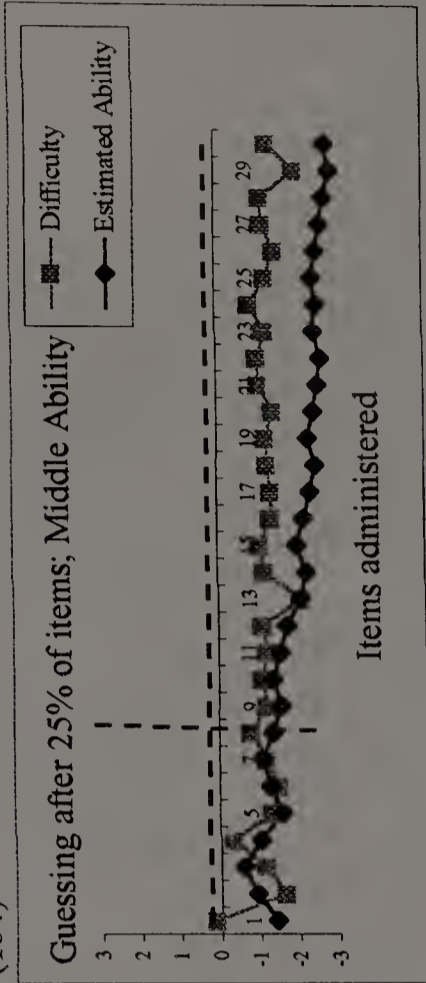
(162)



(163)

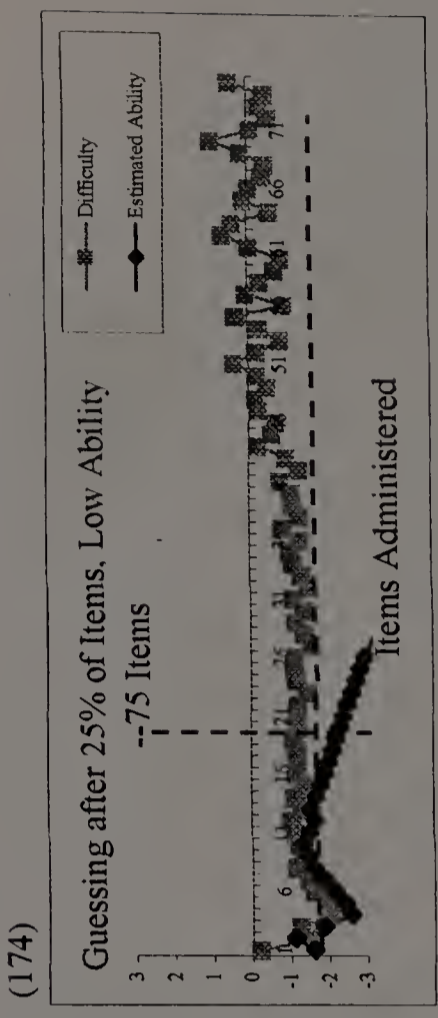
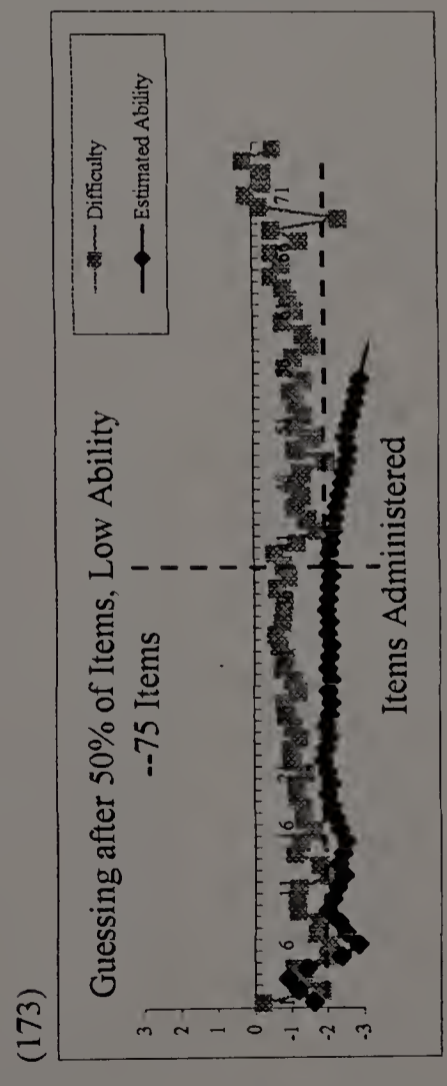
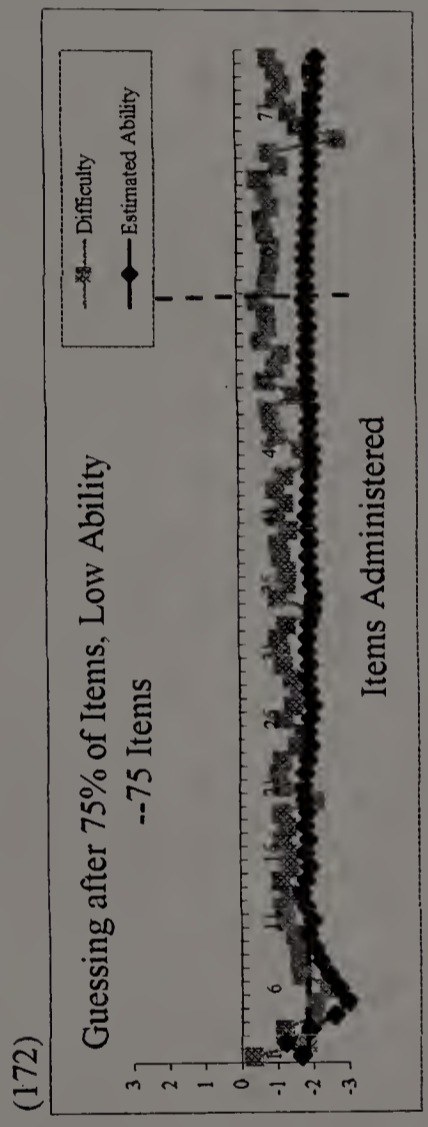
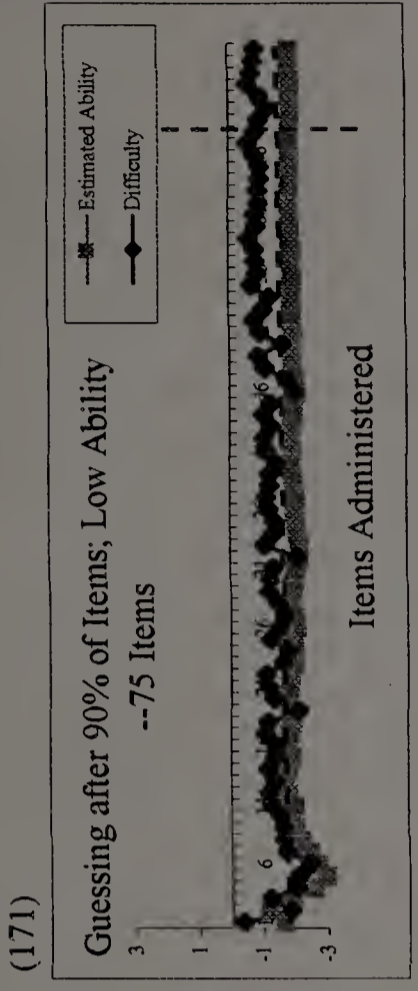
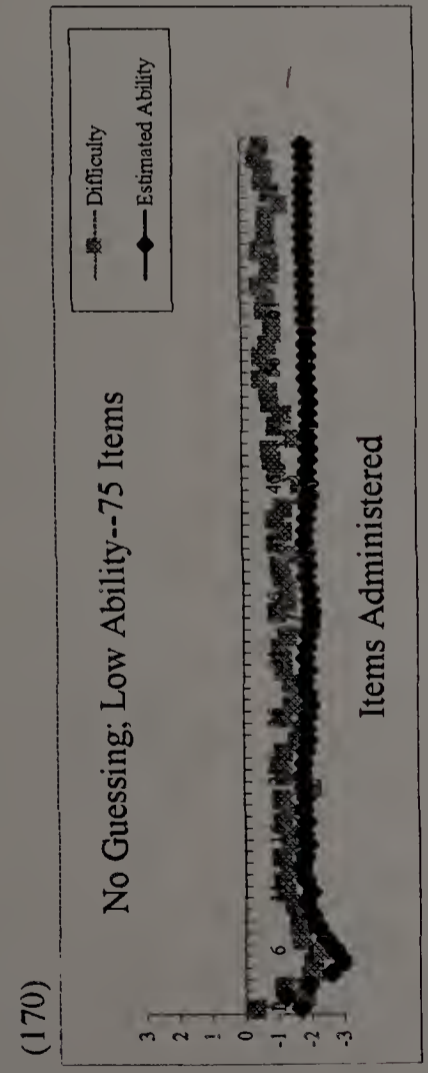


(164)



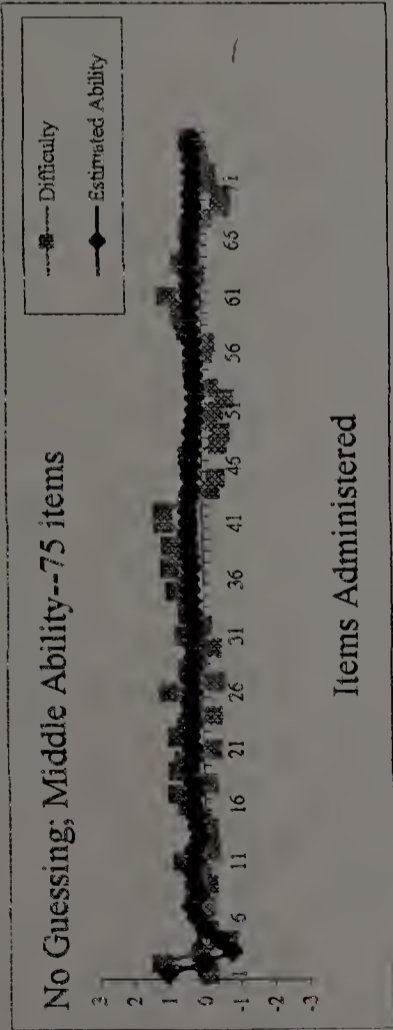


CAT Administration for a Low Ability Examinee who Guesses at a Certain Point in the Test (ARE--75 Items)

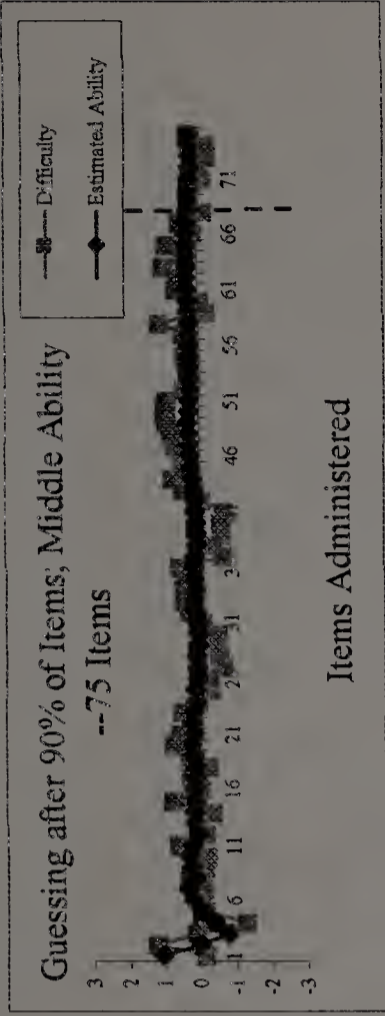


CAT Administration for a Middle Ability Examinee who Guesses at a Certain Point in the Test (ARE--75 Items)

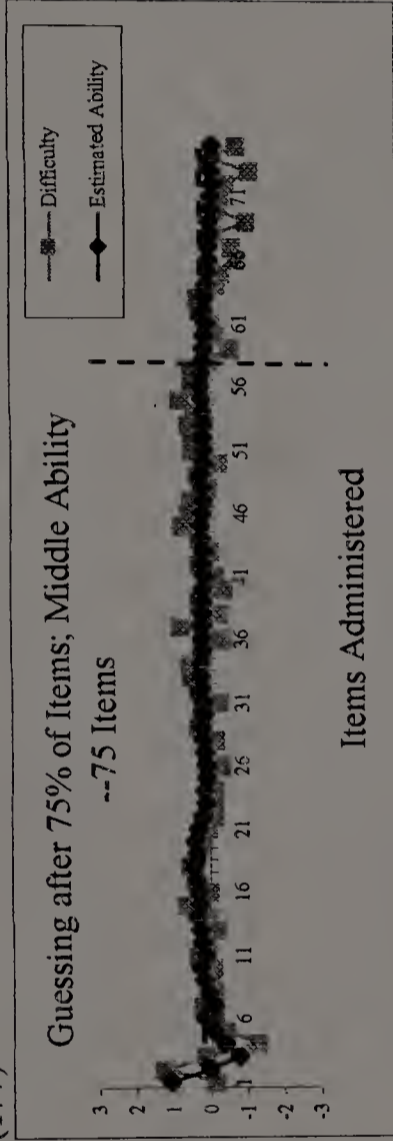
(175)



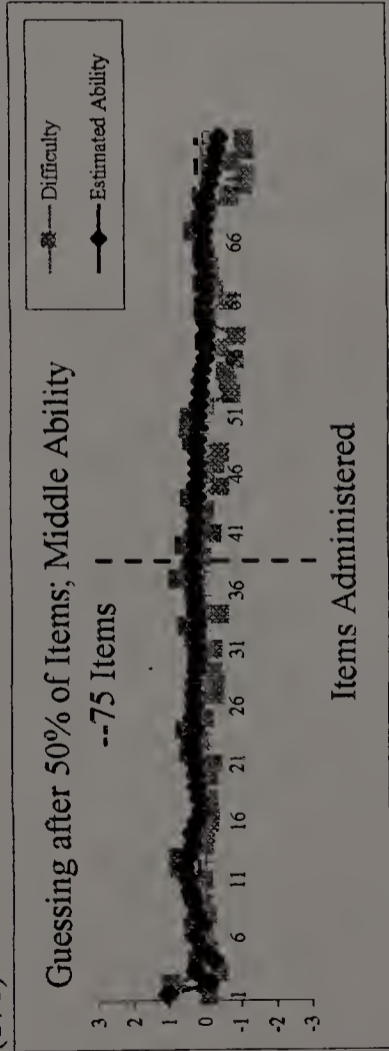
(176)



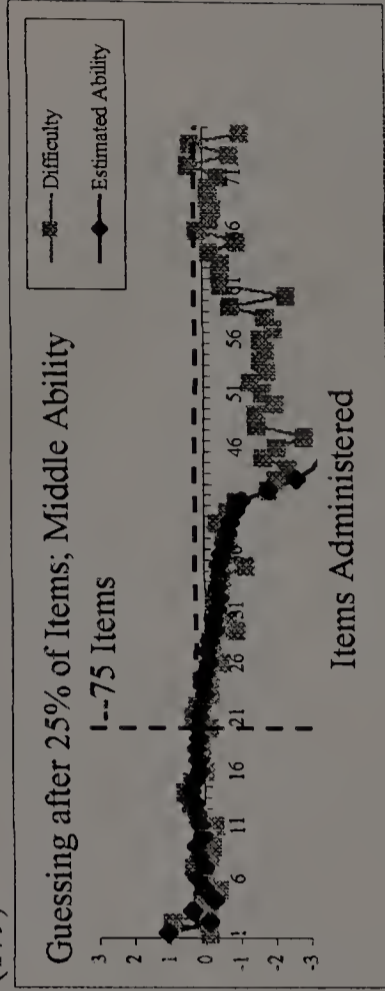
(177)



(178)

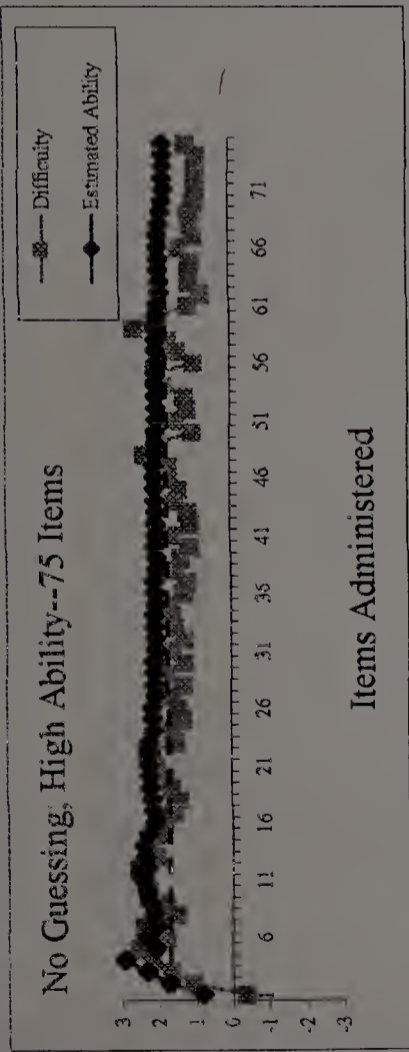


(179)

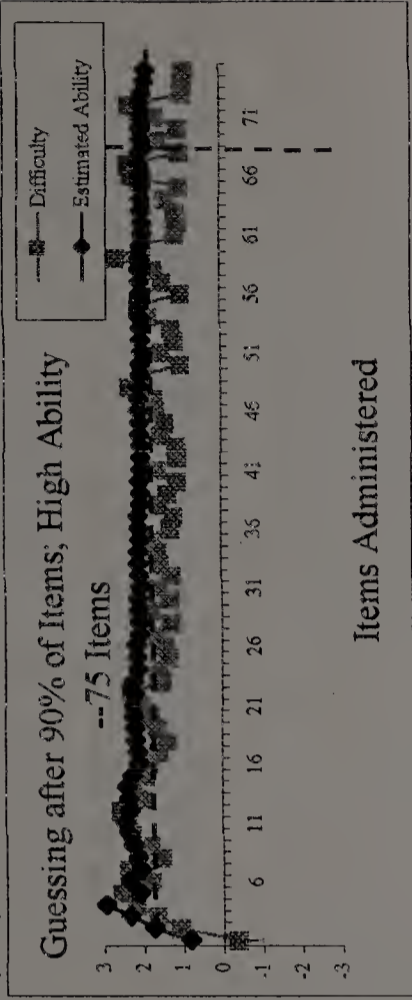


CAT Administration for a High Ability Examinee who Guesses at a Certain Point in the Test (ARE--75 Items)

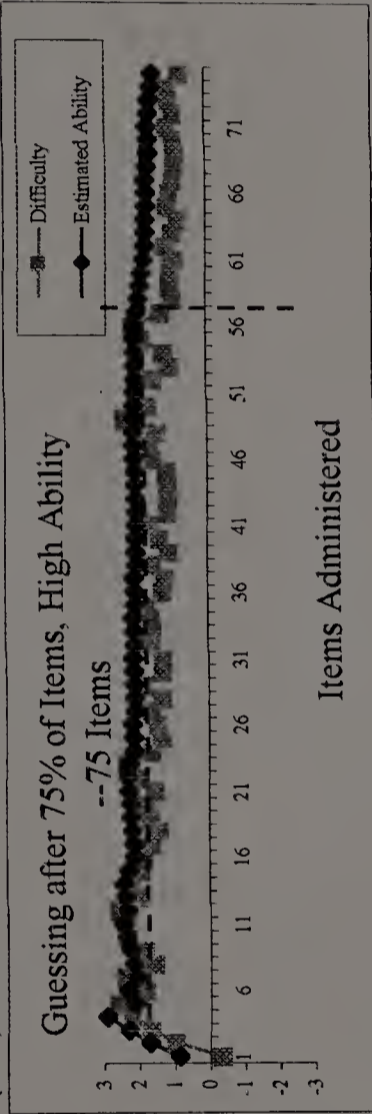
(180)



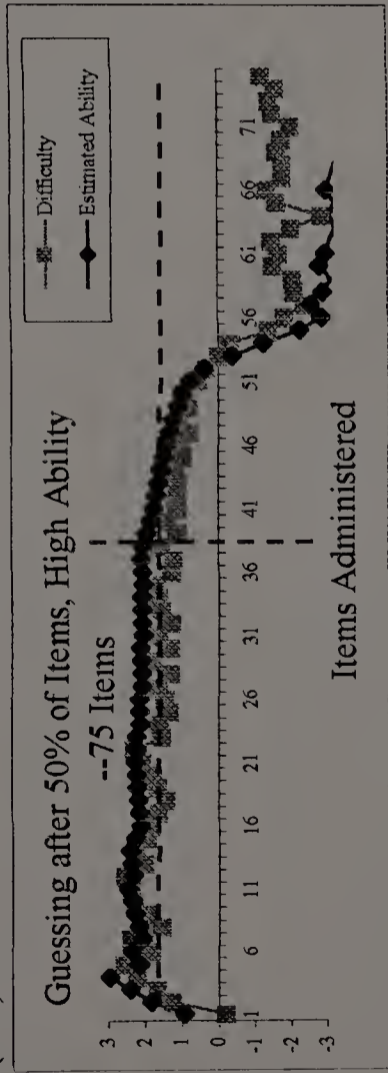
(181)



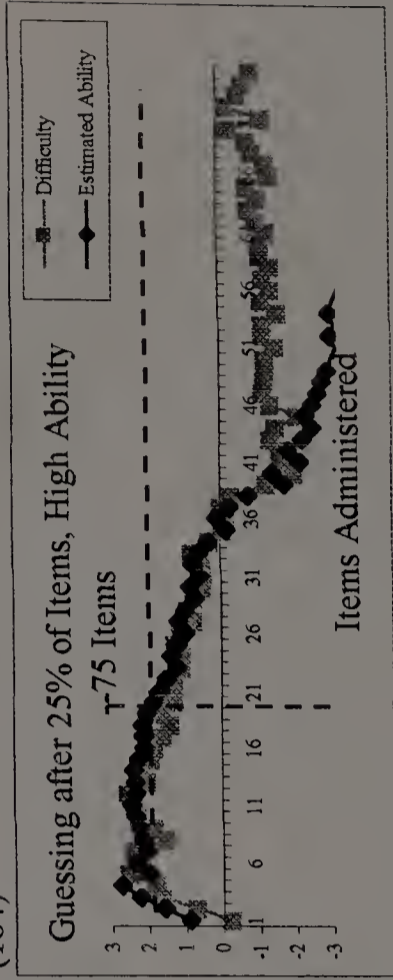
(182)



(183)

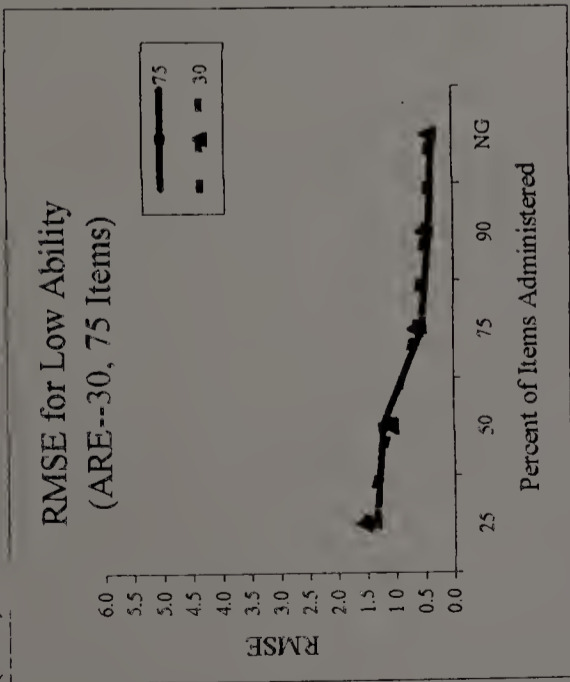


(184)

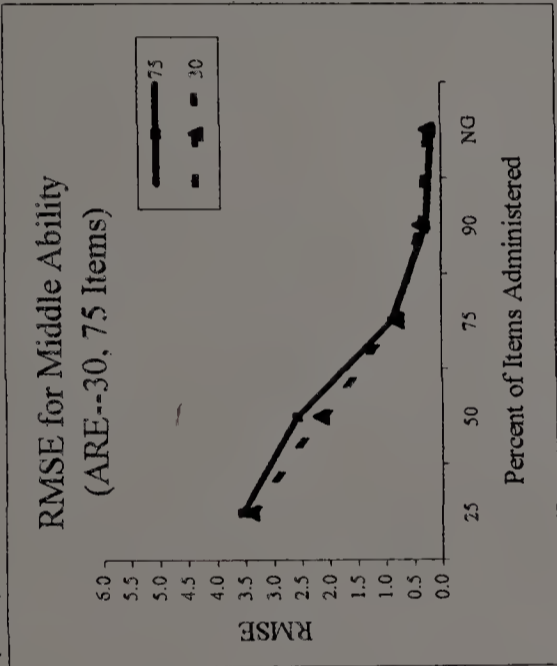


RMSE of Estimates around True Ability for 5 Guessing Scenarios  
 (Performance Testing with AICPA Parameters for 30 and 75 Items on ARE)

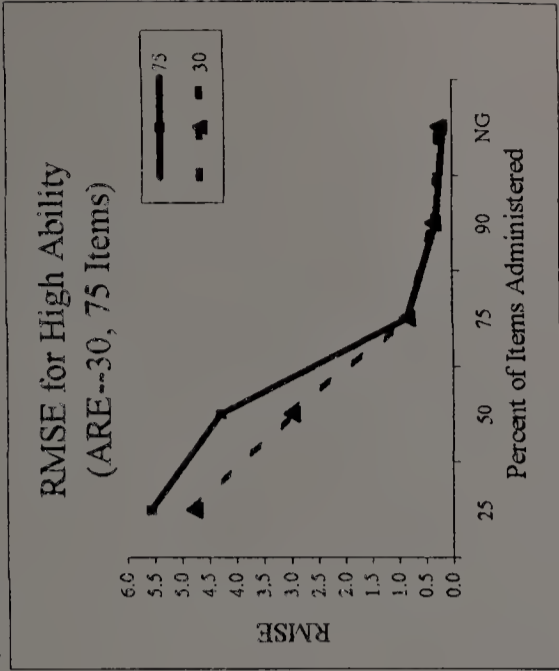
(185)



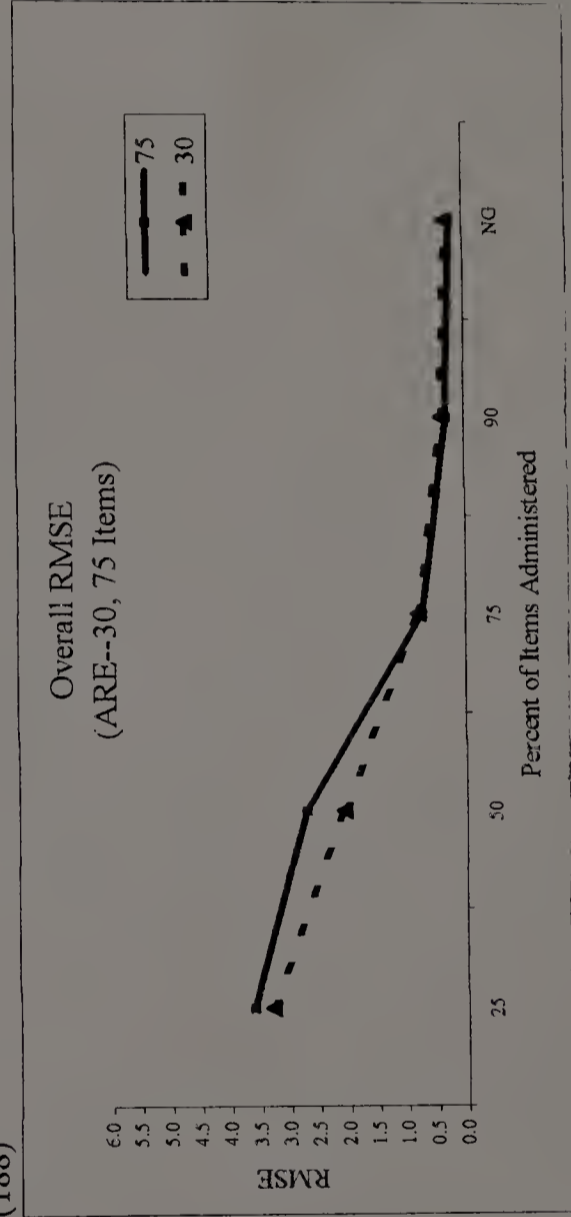
(186)



(187)

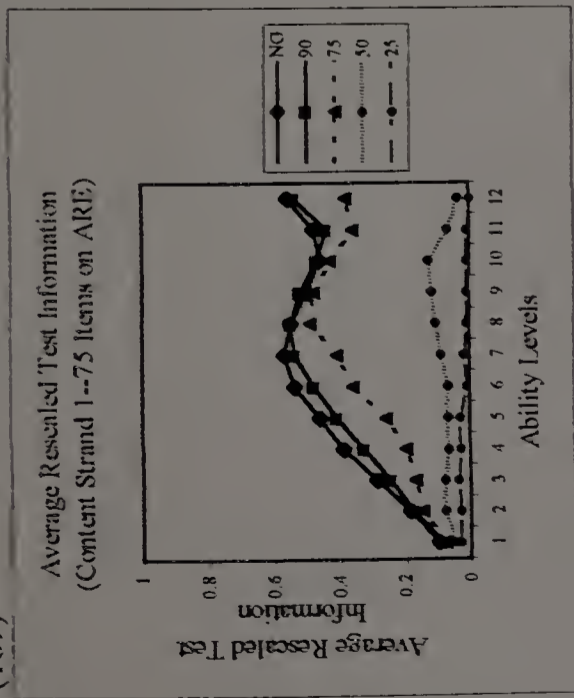


(188)

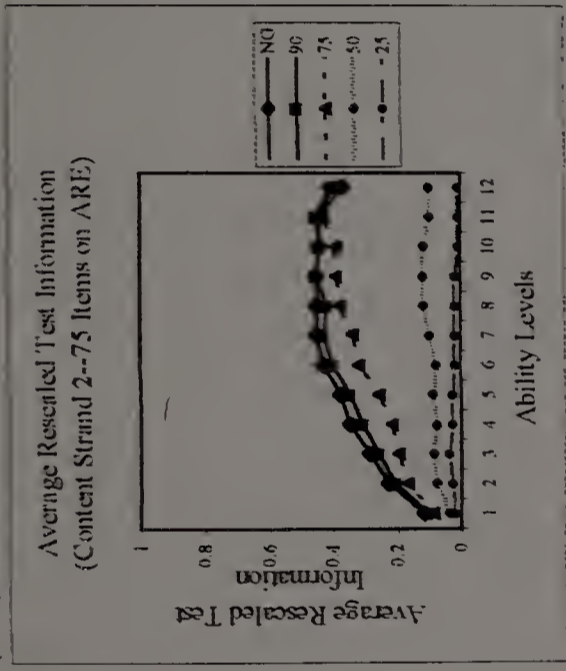


Average Rescaled Test Information in each Content Area at each Ability Level  
 (Performance Testing with AICPA Parameters for 75 Items on ARE)

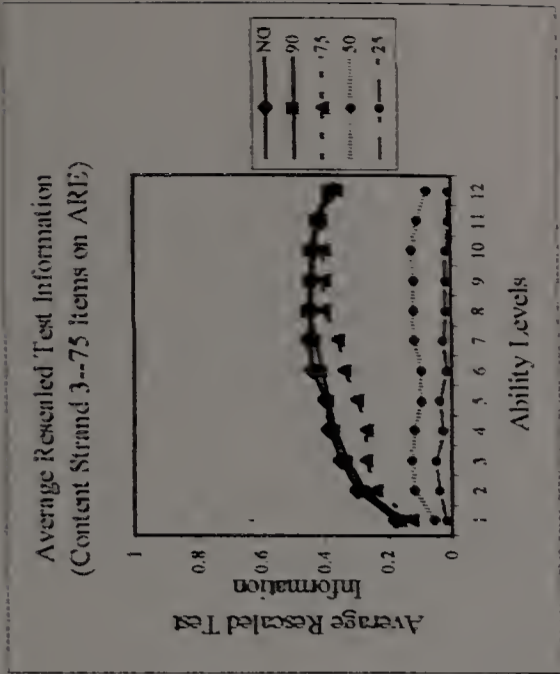
(189)



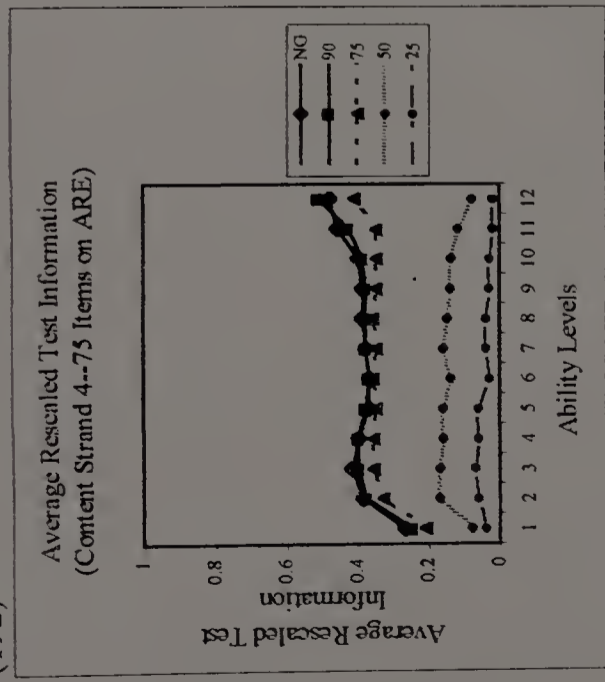
(190)



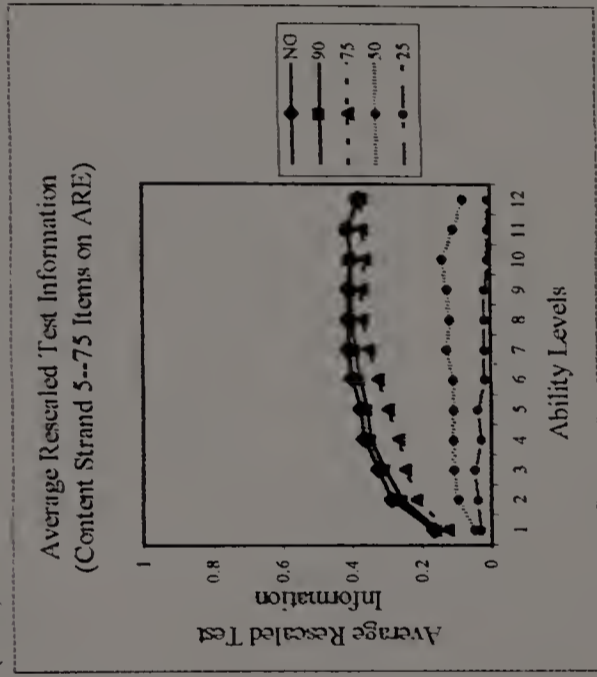
(191)



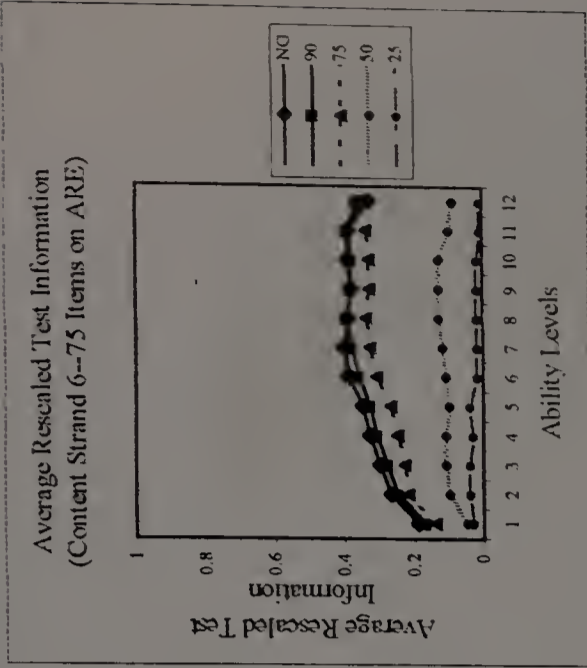
(192)



(193)

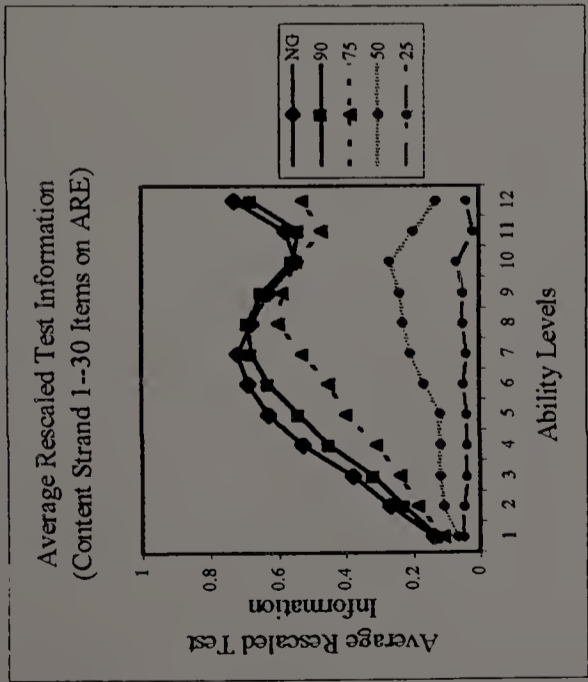


(194)

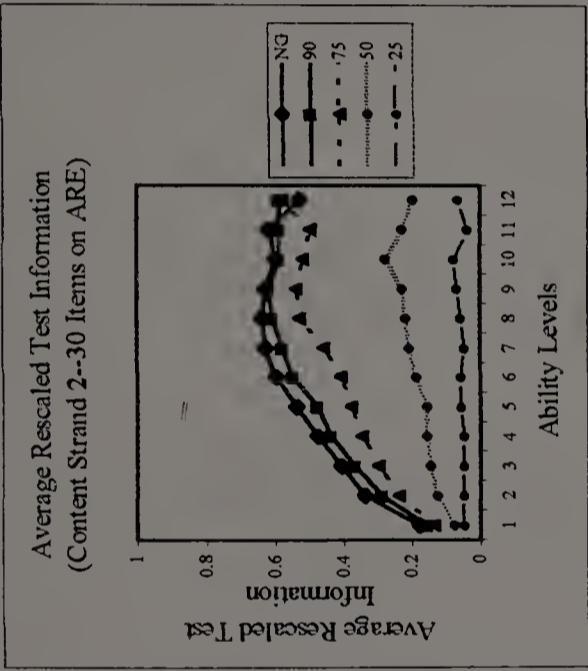


Average Rescaled Test Information in each Content Area at each Ability Level  
 (Performance Testing with AICPA Parameters for 30 Items on ARE)

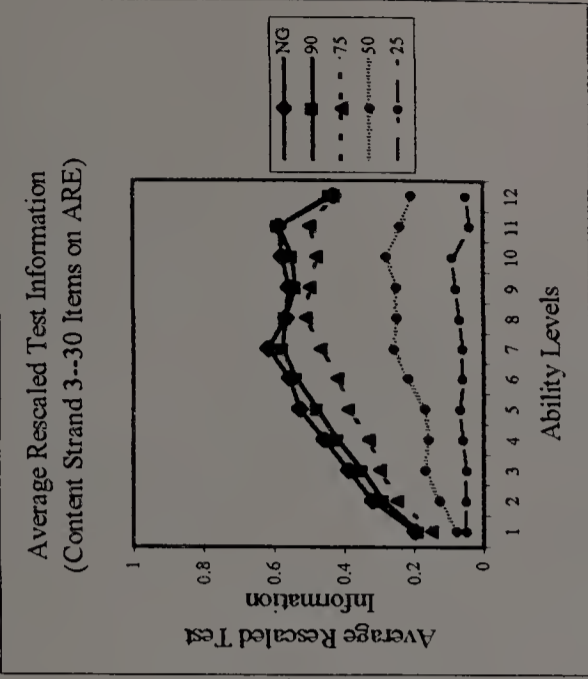
(195)



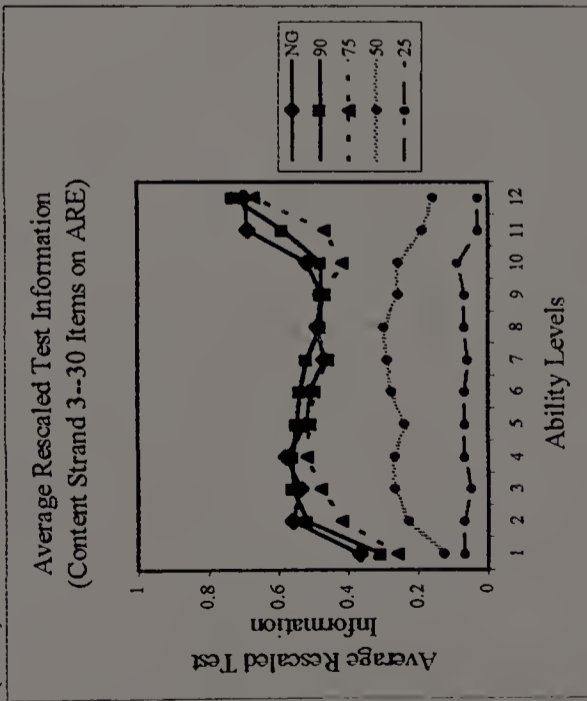
(196)



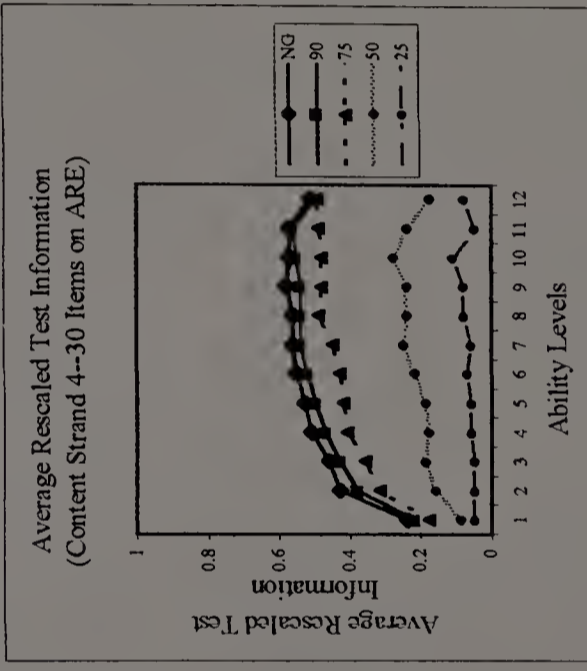
(197)



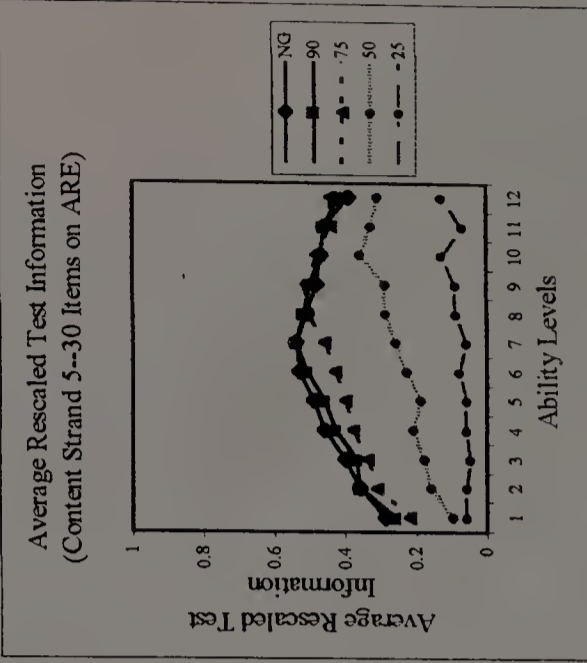
(198)



(199)

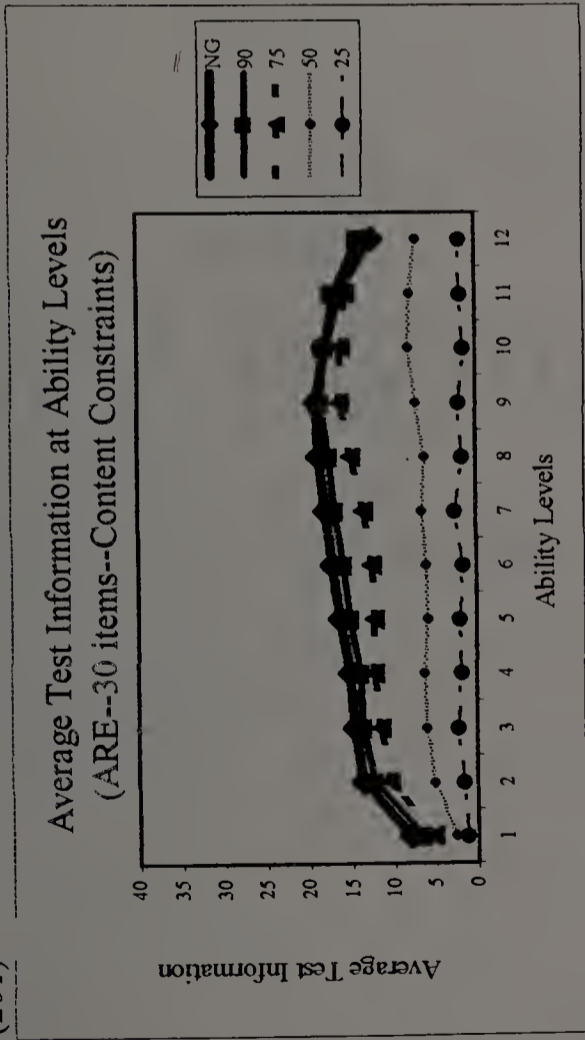


(200)

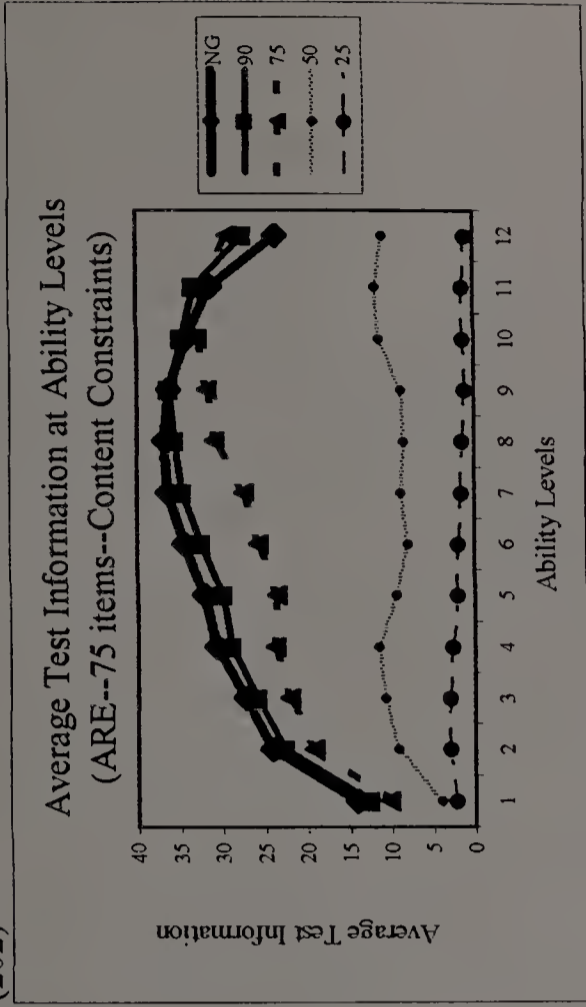


Average Test and Pool Information at each Ability Level for 5 Guessing Scenarios  
 (Performance Testing with AICPA Parameters for 30 and 75 Items on ARE with Content Constraints)

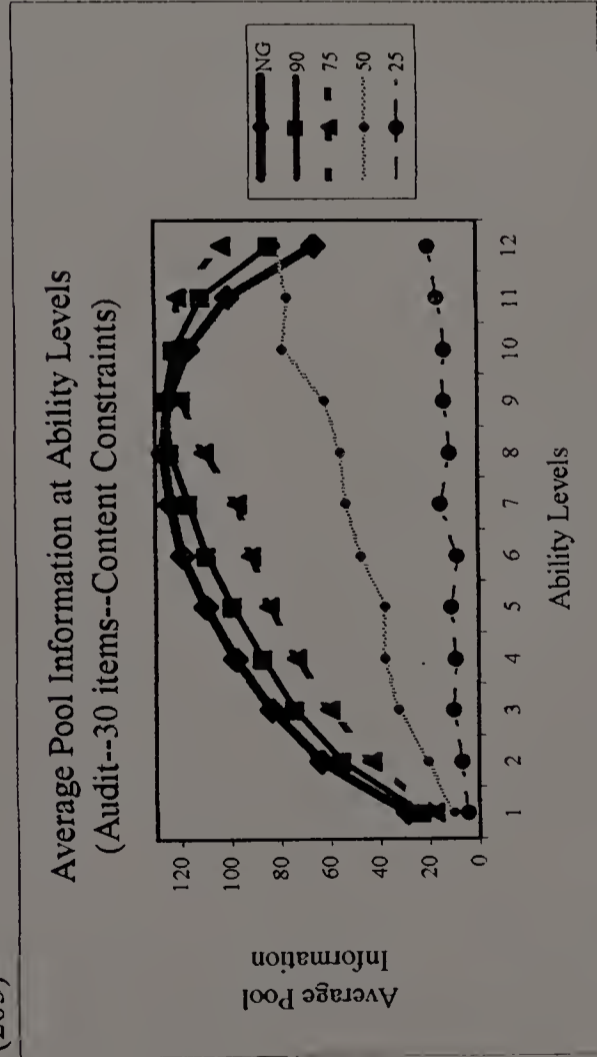
(201)



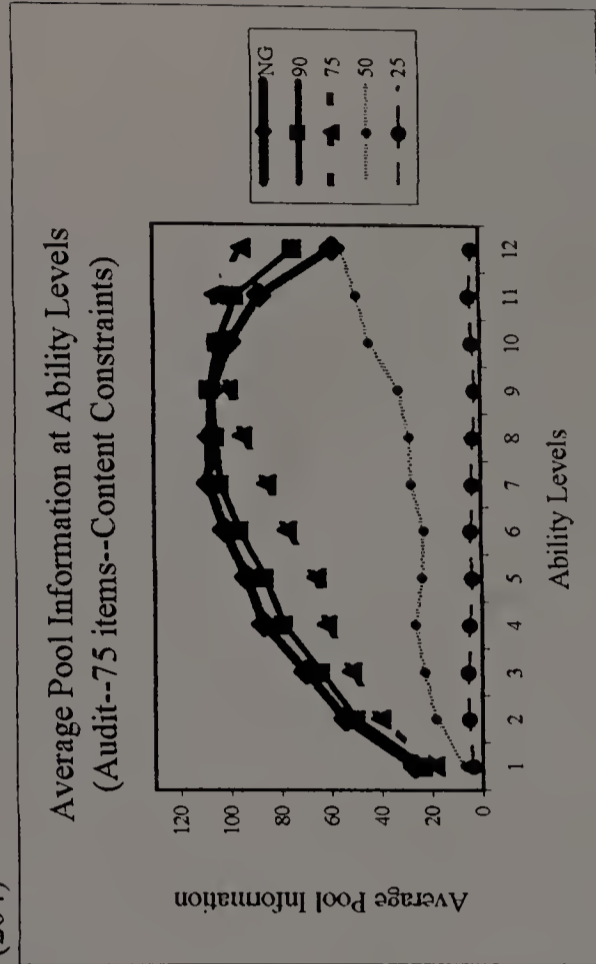
(202)



(203)

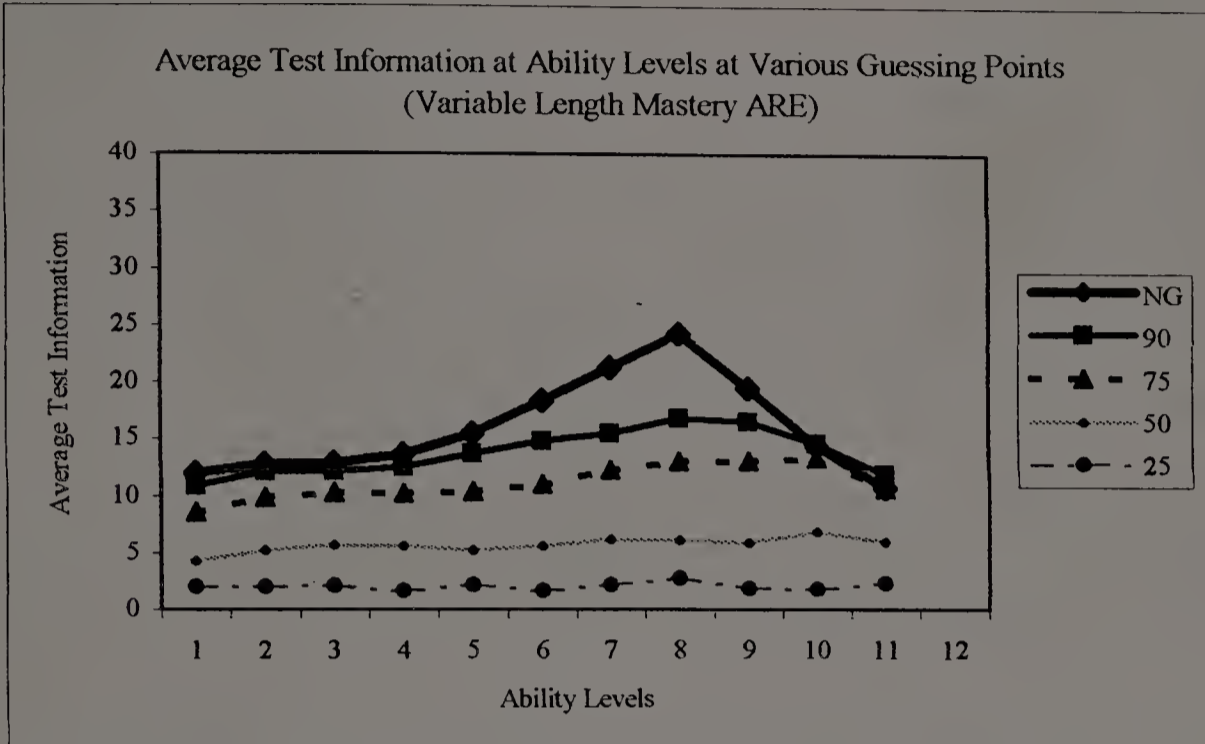


(204)

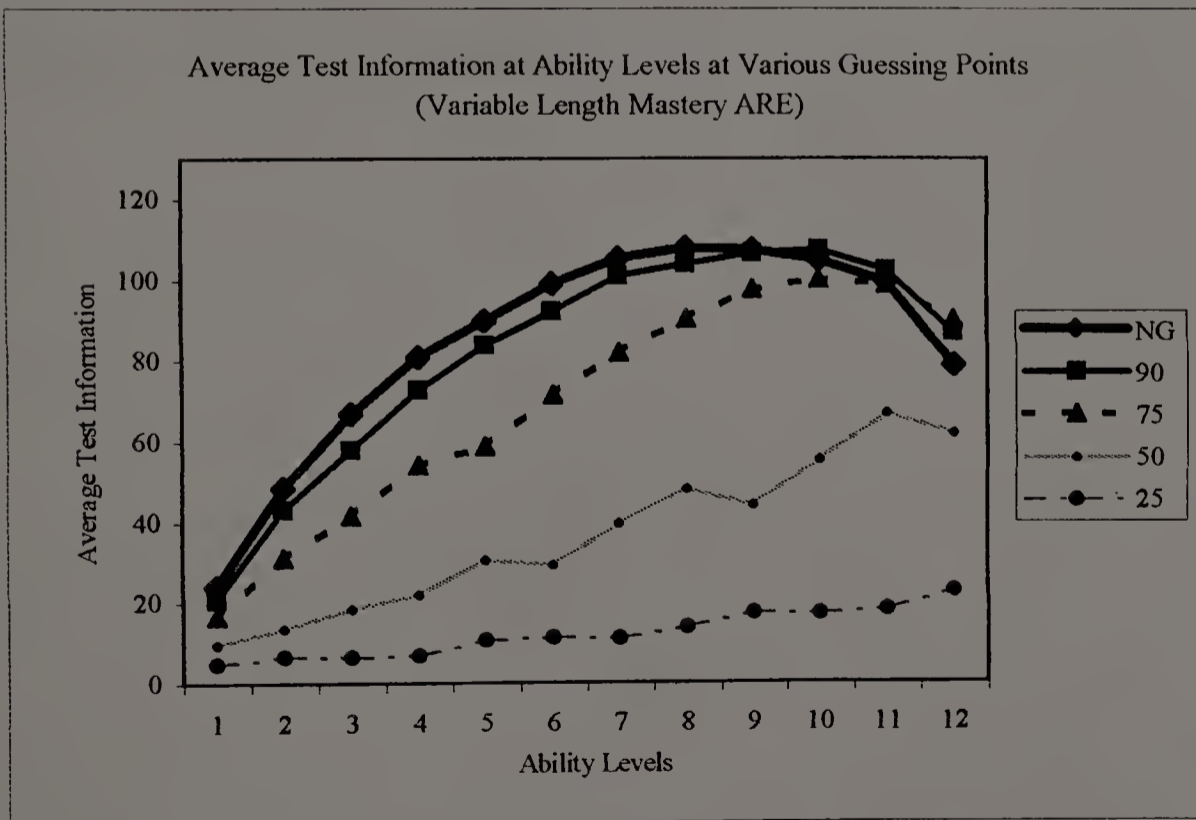


Average Test and Pool Information at each Ability Level for 5 Guessing Scenarios  
 (Variable Length Mastery Testing with AICPA Parameters -- ARE)

(205)

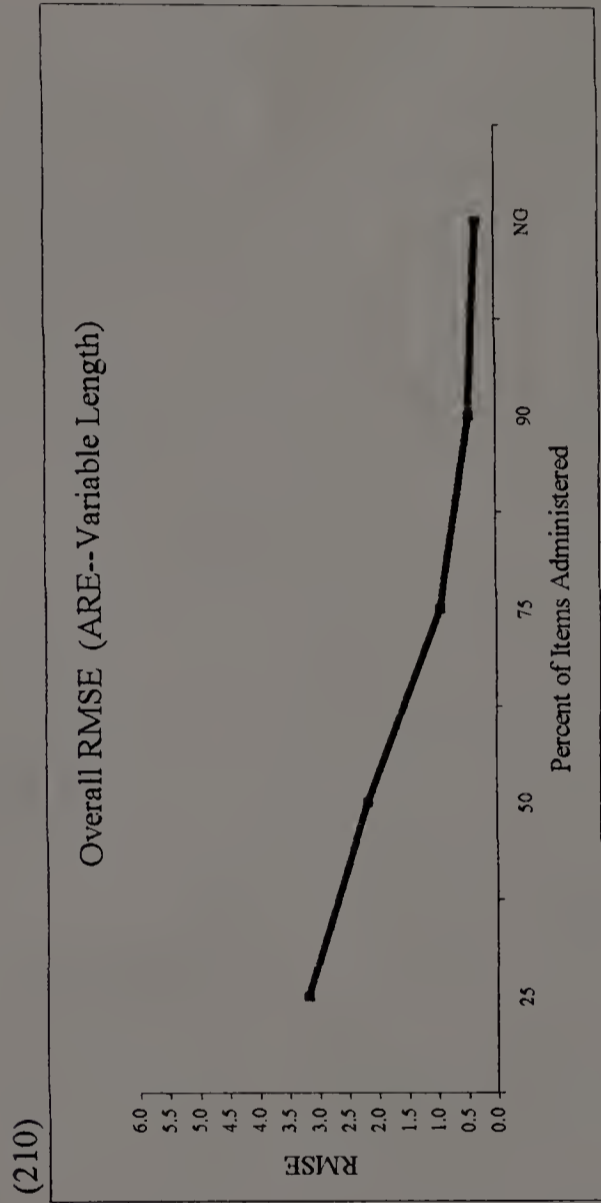
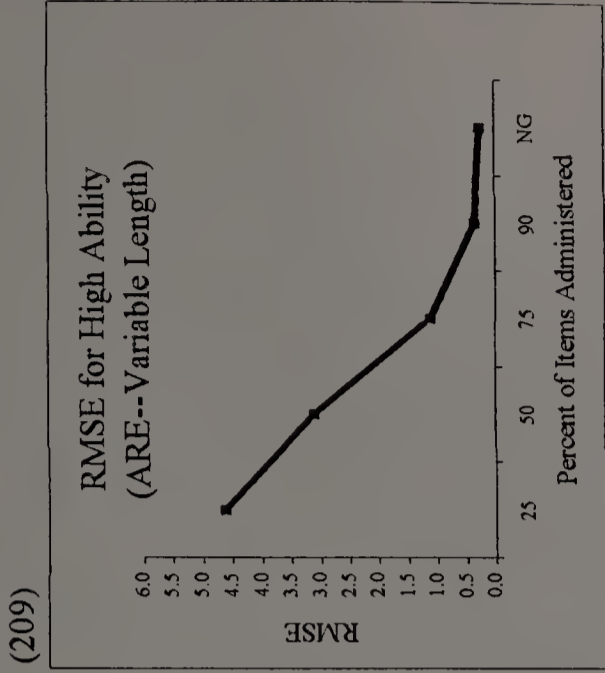
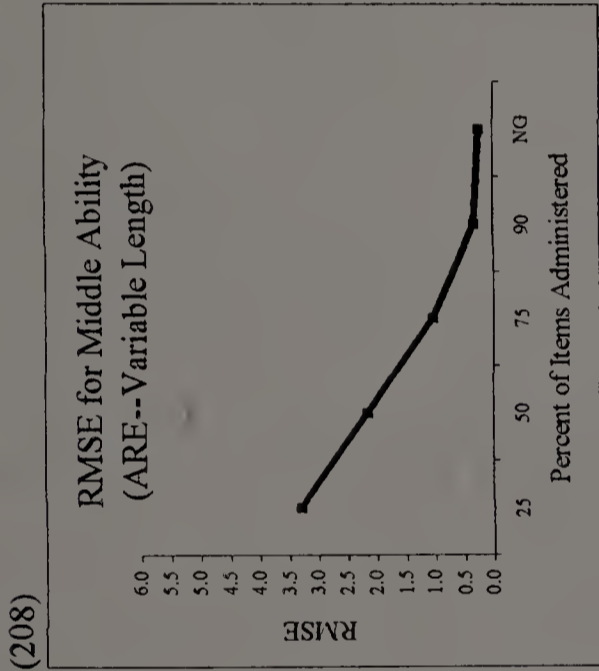
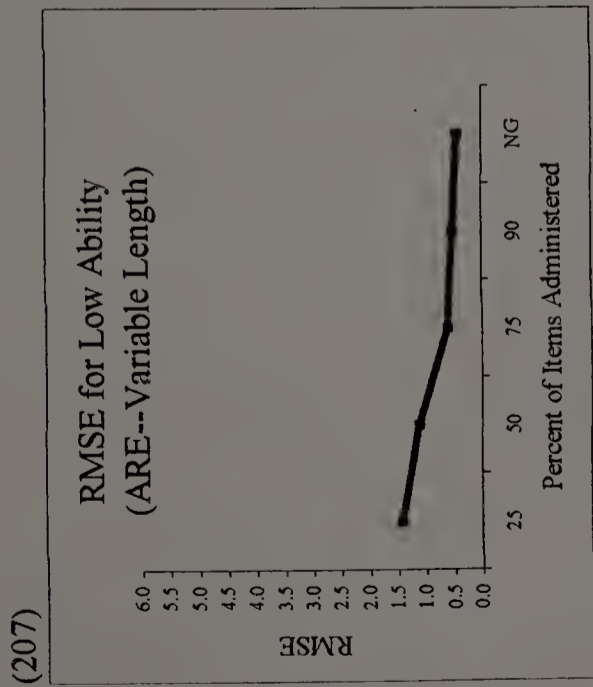


(206)



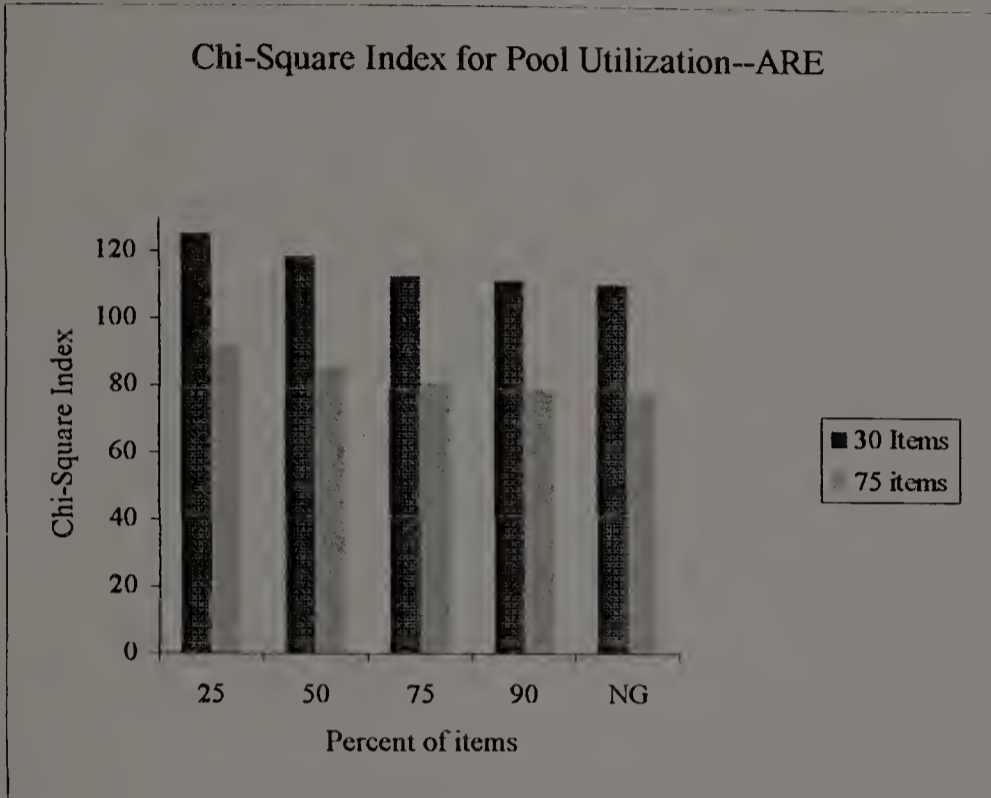


RMSE of Estimates around True Ability for 5 Guessing Scenarios  
 (Variable Length Mastery Testing with AICPA Parameters on ARE with Content Constraints)



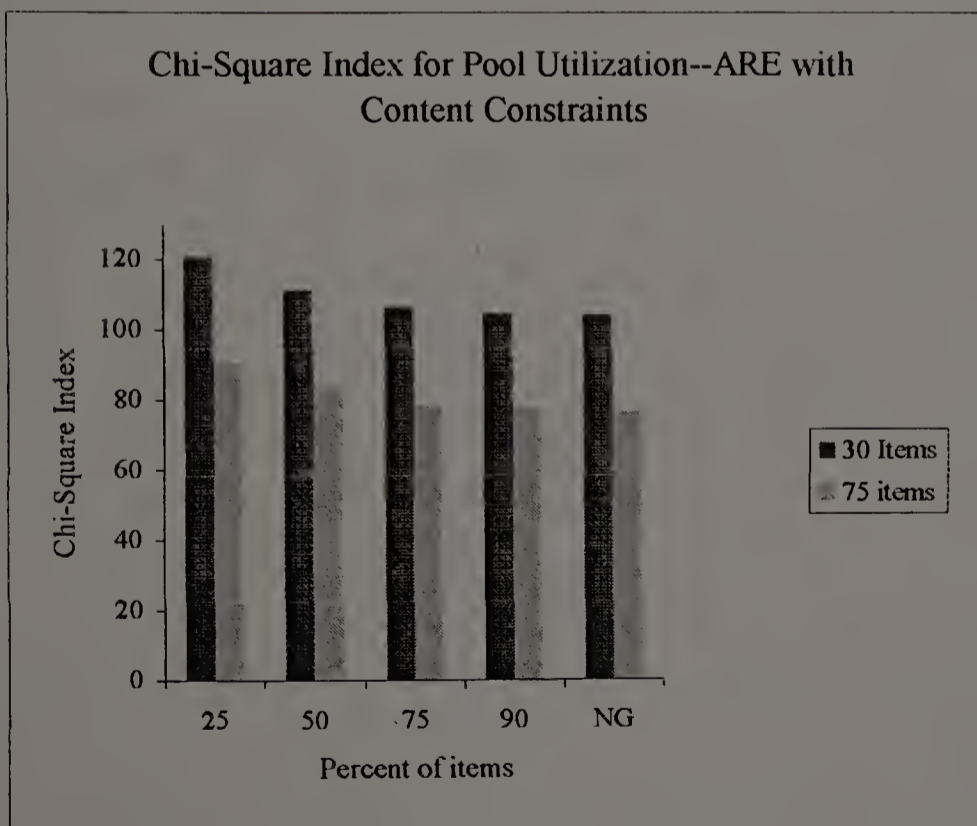
Pool Utilization Index at each Ability Level for 5 Guessing Scenarios  
(Performance Testing with AICPA Parameters for 30 and 75 items on ARE  
without/with Content Constraints)

(211)



	30 items	75 items
25	125.07	91.43
50	118.54	85.60
75	112.89	81.10
90	111.13	79.07
NG	110.27	78.10

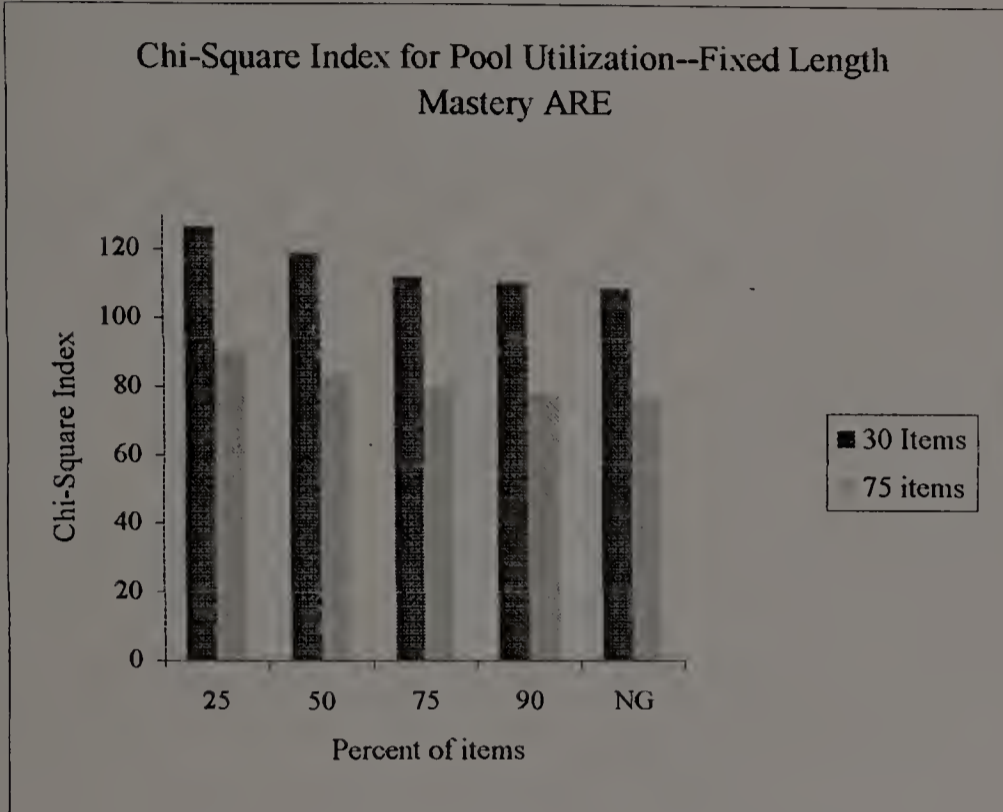
(212)



	30 items	75 items
25	120.81	90.33
50	111.39	84.03
75	106.36	78.74
90	104.60	77.37
NG	104.17	76.66

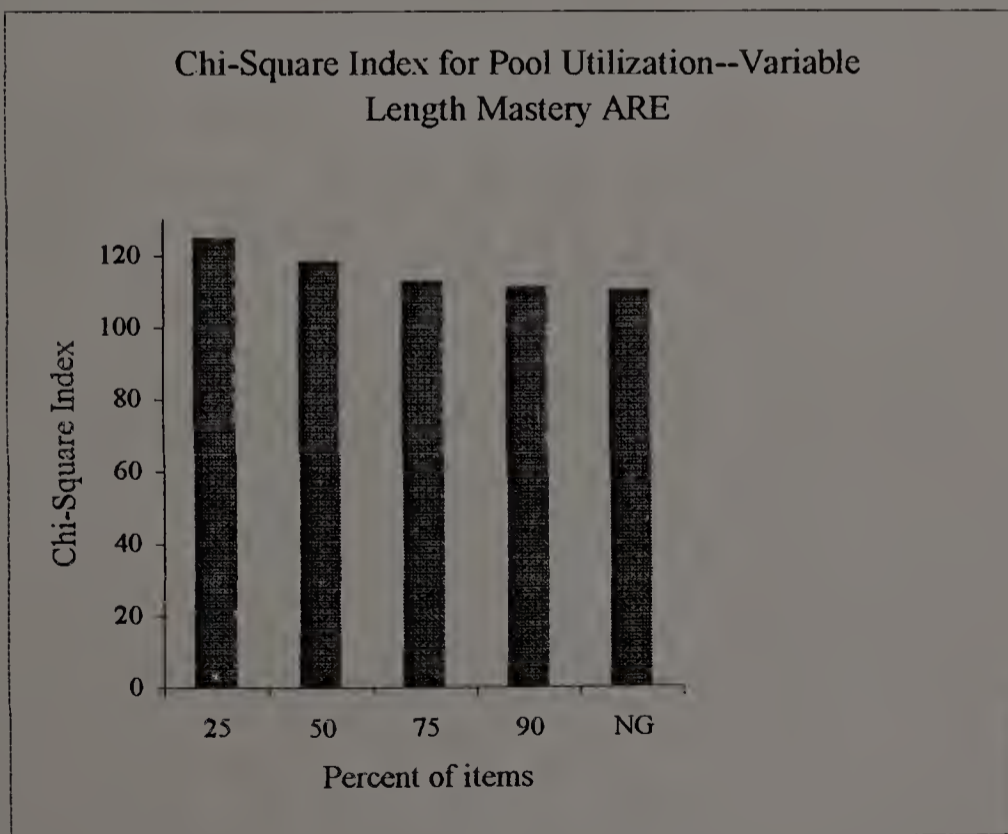
Pool Utilization Index at each Ability Level for 5 Guessing Scenarios  
(Mastery Testing with AICPA Parameters for Fixed/Variable Length ARE)

(213)



	30 items	75 items
25	125.07	91.43
50	118.54	85.60
75	112.89	81.10
90	111.13	79.07
NG	110.27	78.10

(214)



	Variable Length
25	128.49
50	123.67
75	119.74
90	119.20
NG	114.79

## BIBLIOGRAPHY

- American Council on Education (1995). Guidelines for computerized adaptive test development and Use in Education. Washington DC: Author.
- Ackerman, T. (April, 1987). The use of unidimensional item parameter estimates of multidimensional items in adaptive testing. Paper presented at the annual meeting of the American Educational Research Association Conference, Washington DC. (ERIC Document Reproduction Service No. Ed 284 901)
- Bergstrom, B. & Garshon, R. (April, 1994). Computerized adaptive testing exploring examinee response time using hierarchical linear modeling. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA. (ERIC Document Reproduction Service No. Ed 400 286)
- Bergstrom, B. & Garshon, R. (April, 1992). Comparison of item targeting strategies for pass/fail computer adaptive tests. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA. (ERIC Document Reproduction Service No. Ed 400 287)
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Ed.), Statistical theories of mental scores (chapters 17-20). Reading, Massachusetts: Addison-Wesley.
- Bontempo, B. D. & Julian, E. R. (March, 1997). Paper presented at the annual meeting of the National Council on Measurement in Education. Chicago, IL.
- Bracey, G., Rudner, L. M. (1992). Person-Fit Statistics: High potential and many unanswered questions. Research Report 92-5. Office of Educational Research and Improvement (ED), Washington, DC.
- Bradlow, E. T., Weiss, R. E. & Cho, M. (1998). Bayesian identification of outliers in computerized adaptive tests. Journal of the American Statistical Association, 94(443), 910-919.
- Carlson, R. (1994). Computer Adaptive Testing: a shift in the evaluation paradigm. Educational Technology Systems, 22(3), 213-224.
- Chang, H. (April, 1996). A model for score maximization within a computerized adaptive testing environment. Paper presented at the annual meeting of the National Council on Measurement in Education. New York, NY.

- Chang, H., Qian, J., & Ying, Z. (April, 2000).  $\alpha$ -Stratified multistage CAT with b-blocking. Paper presented at the annual meeting of the National Council on Measurement in Education. New Orleans, LA.
- Chang, H. & Ying, Z. (March, 1997). A global information approach to computerized adaptive testing. Paper presented at the annual meeting of the National Council on Measurement in Education. Chicago, IL.
- Chang, H. & Ying, Z. (1999).  $\alpha$ -Stratified multistage computerized testing. Applied Psychological Measurement, 23(3), 211-222.
- Clark, C. (1976). Proceedings of the first conference on computerized adaptive testing. Washington DC: US Government Printing Office.
- Dodd, B. G. (1990). The effect of item selection procedure and step size on computerized adaptive attitude measurement using the rating scale model. Applied Psychological Measurement, 14(4), 355-366.
- Drasgow, F. & Parsons, C. (1983). Application of unidimensional item response theory to multidimensional data. Applied Psychological Measurement, 7, 218-232.
- Eggen, T. H. (1999). Item selection in adaptive testing with the sequential probability ratio test. Applied Psychological Measurement, 23(3), 195-210.
- Eignor, D. R., Stocking, M. L., Way, W. D. & Steffen, M. Case studies in computer adaptive design through simulation. Research Report RR-93-56. Educational Testing Service, Princeton, NJ.
- Glas, C. A., Meijer, R. R. & Van Krimpen-Stoop, E. (1998). Statistical Tests for Person Misfit in Computerized Adaptive Testing. Research Report 98-01. Educational Science and Technology, University of Twente, Netherlands.
- Green, B., Bock, R., Humphreys, L. & Reckase, M. (1984). Technical guidelines for assessing computerized adaptive tests. Journal of Educational Measurement, 21(4), 347-360.
- Guttman, L. A. (1944). A basis for scaling qualitative data, as described in Weiss, D. J. (1983).
- Hambleton, R. & Swaminathan, H. (1985). Item response theory: principles and applications. Massachusetts: Kluwer Academic Publishers.
- Hambleton, R., Swaminathan, H. & Rogers, J. (1991). Fundamentals of item response Theory. Newbury: Sage Publications.

- Hambleton, R. (1983). Applications of Item Response Theory. Vancouver: Educational Institute of British Columbia.
- Hambleton, R. K. & Pieters, P. M. & Zaal, N. J. (1991). Computerized adaptive testing: theory, applications, and standards. In Hambleton, R. K. & Zaal, N. J. (Ed.). Advances in educational and psychological testing: theory and applications (pp. 341-366). Massachusetts: Kluwer Academic Publishers.
- Hick, W. E. (1951). Information theory and intelligence tests, as described in Kreitzberg et al. (1978).
- Jensema, C. J. (1974). The Validity of Bayesian Tailored Testing. Educational & Psychological Measurement, 34(4), 757-66.
- Kalohn, J. & Spray, J. (1998). Effect of item selection on item exposure rates within a computerized classification test. Paper presented at the annual meeting of the National Council on Measurement in Education. San Diego, CA.
- Kingsbury, G. & Houser, R. (1993). Assessing the utility of Item Response Models: Computerized Adaptive Testing. Educational Measurement: Issues and Practice, 2(12), 21-26.
- Koch, W. & Reckase, M. (1979). Problems in application of latent trait models to tailored testing. Paper presented at the annual meeting of the National Council on Measurement in Education. San Francisco, CA. (ERIC Document Reproduction Service No. Ed 177 196)
- Kogut, Jan. (1986). A review of IRT-Based Indices for Detecting and Diagnosing Aberrant Response Patterns. Report 86-4. Department of Education, Twente University, Netherlands.
- Kogut, Jan. (1987). Detecting aberrant response patterns in the Rasch Model. Research Report 87-3. Department of Education, Twente University, Netherlands.
- Kogut, Jan. (1987). Reduction of bias in Rasch estimates due to Aberrant Patterns. Research Report 87-5. Department of Education, Twente University, Netherlands.
- Kreitzberg, C., Stocking, M. & Swanson, L. (1978). Computerized adaptive testing: principles and directions. Computers & Education, 2, 319-329.
- Llabre M. & Froman, T. W. (1987). Allocation of Time to Test Items: A Study of Ethnic Differences. Journal of Experimental Education, 55(3) 137-140.
- Laurier, M. (April, 1990). What can we do with computerized adaptive testing and what we cannot do?. Paper presented at the annual meeting of the Regional Language

Center Seminar, Singapore. (ERIC Document Reproduction Service No. Ed 322 729)

Lawley, D. N. (1943). On problems connected with item selection and test construction, as described in Weiss, D. J. (1983).

Leucht, R. M., Nungester, R. J. & Hadadi, A. (April, 1996). Heuristics based CAT: Balancing item information, content and exposure. Paper presented at the annual meeting of the National Council on Measurement in Education. New York, NY.

Lord, F. M. (1952). A theory of test scores. Psychometric monograph. No. 7.

Lord, F. M. & Novick, M. R. (1968). Statistical theories of mental scores. Massachusetts: Addison-Wesley.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. New Jersey: Hillsdale.

McBride, J. & Wiess, D. (1983). Bias and Information of Bayesian Adaptive Testing (Research Report 83-2). Minnesota: Minnesota University.

McLeod L. D. & Lewis, C. (April, 1996). Person-Fit indices and their role in the CAT environment. Paper presented at the annual meeting of the National Council on Measurement in Education. New York, NY.

Meijer, R. R. (1994). The influence of the presence of deviant score patterns on the power of a person-fit statistic. Research Report 94-1. Educational Science and Technology, University of Twente, Netherlands.

Meijer, R. R. & Sijsma, K. (1994). Detection of aberrant score patterns: a review of recent developments. Research Report 94-8. Educational Science and Technology, University of Twente, Netherlands.

Mills, C. & Stocking, M. (1996). Practical issues in large scale high stakes computerized adaptive testing. Applied Measurement in Education, 9(4), 287-304.

Mosier, C. I. (1941). Psychophysics and mental test scores: the constant process, as described in Weiss, D. J. (1983).

Nering, M. L. (1995). The Distribution of Person-fit Using True and Estimated Person Parameters. Applied Psychological Measurement, 19(2) 121-29.

Parshall, C. G., Kromrey, J. D. & Hogarty, K. Y. (April, 2000). Sufficient simplicity or comprehensiveness complexity? A comparison of probabilistic and stratification methods

of exposure control. Paper presented at the annual meeting of the National Council on Measurement in Education. New Orleans, LA.

Reise, S. P., Due, A. M. (1991). The influence of test characteristics on the detection of aberrant response patterns. Applied Psychological Measurement, 15(3), 217-226.

Revuelta J. & Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. Journal of Educational Measurement, 35(4) 311-327.

Robin, F. (2000). Computer Based Testing Software (CBTS). Laboratory of psychometric and evaluative research. Amherst: University of Massachusetts.

Schartz, M. (April, 1986). Measuring up in an individualized way with CAT-ASVAB: considerations in the development of adaptive testing pools. Paper presented at the annual meeting of the American Educational Research Association Conference. San Francisco, CA. (ERIC Document Reproduction Service No. Ed 269 463)

Schaefer, G., Steffen, M., Golub-Smith, M., Mills, C. & Durso, R. (1995). The introduction and comparability of the computer adaptive GRE General test (Research Report 88-08aP). New Jersey: Educational Testing Service.

Schnipke, Deborah L. (April, 1995). Assessing Speededness in Computer-Based Tests Using Item Response Times. Paper presented at the annual meeting of the National Council on Measurement in Education.

Schnipke, D. L. & Pahley, P. J. (April, 1997). Assessing subgroups differences in item response times. Paper presented at the annual meeting of the National Council on Measurement in Education. Chicago, IL. (ERIC Document Reproduction Service No. Ed 409 364)

Steffen, M, & Way, W. D. (April, 1999). Test-taking strategies in computerized adaptive testing. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Quebec.

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

Stocking, M. (1996). An alternative method for scoring adaptive tests. Journal of Behavioral statistics, 21(4), 365-389.

Stocking, M. & Swanson, L. (1993). A Method for Severely Constrained Item Selection in Adaptive Testing. Applied Psychological Measurement, 17(3) 277-292.



- Sympson, J. B. & Hetter, R. D. (1985). Controlling item exposure rates in computerized adaptive testing, as described in Stocking, M. & Lewis, C. (1996).
- Thissen, D. (1990). Reliability and measurement precision in computerized testing. In Wainer, H. (Ed.). Computerized adaptive testing: A primer (pp. 161-185). New Jersey: Lawrence Erlbaum. Associates.
- Traub, R. E. (1983). A priori considerations in choosing an item response model. In Hambleton, R. (Ed.). Applications of item Response theory (pp. 57-70). Vancouver: Educational Institute of British Columbia.
- Traub, R. & Wolfe, R. (1981). Latent trait theories and the assessment of educational achievement. In Berhner, D. (Ed.). Review of Research in Education, 9, 377-435.
- Wainer, H. (1983). Are we correcting for guessing in the wrong direction?. In Wiess, D. (1983). New horizons in testing (pp. 63-70). New York: Academic Press, Inc.
- Wainer, H. (1990). Computerized adaptive testing: A primer. New Jersey: Lawrence Erlbaum Associates.
- Wang, W., Wilson, M., & Adams. (1995). Item response modeling for multidimensional between-items and multidimensional within-items. Paper presented at the International Objective Measurement Conference. Berkeley, CA.
- Weiss, D. (1982). Improving measurement quality and efficiency with adaptive testing. Applied psychological Measurement, 6, 473-492.
- Wiess, D. (1983). New horizons in testing. New York: Academic Press, Inc.
- Wiess, D. & Kingsbury, G. (1983). A comparison between IRT-based adaptive mastery testing and sequential mastery testing. In D. Wiess (Ed.). New horizons in testing (pp. 257-283). New York: Academic Press, Inc.
- Wise, S. (March, 1997). Examinees issues in CAT. Paper presented at the annual meeting of the National Council on Measurement in Education. Chicago, IL. (ERIC Document Reproduction Service No. Ed 408 329)
- van der Linden, W. J. (1986). The changing conception of measurement in education and psychology. Applied Educational Measurement, 10(4), 325-332.
- van der Linden, W. J. & Scrams, D. J. & Schnipke, D. L. (1999). Using response time constraints to control for differential speededness in computerized adaptive testing. Applied Psychological Measurement, 23(3), 195-210.

Yi, Q. & Nering, M. L. (1999). Simulating nonmodel-fitting responses in a CAT environment. Paper presented at the annual meeting of the National Council on Measurement in Education. Chicago, IL. (ERIC Document Reproduction Service No. Ed 427 042)

Yoes, M. E. & Ho, K. T. (April, 1991). The degree of Person Misfit on a nationally standardized achievement test. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

