

1-1-1993

Optimal test designs with content balancing and variable target information functions as constraints.

Tit Loong Lam

University of Massachusetts Amherst

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

Recommended Citation

Lam, Tit Loong, "Optimal test designs with content balancing and variable target information functions as constraints." (1993).

Doctoral Dissertations 1896 - February 2014. 4998.

https://scholarworks.umass.edu/dissertations_1/4998

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

UMASS/AMHERST



312066013025843

OPTIMAL TEST DESIGNS WITH CONTENT BALANCING AND VARIABLE
TARGET INFORMATION FUNCTIONS AS CONSTRAINTS

A Dissertation Presented

by

LAM TIT LOONG

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

DOCTOR OF EDUCATION

February 1993

School of Education

© Copyright by Lam Tit Loong 1993
All Rights Reserved

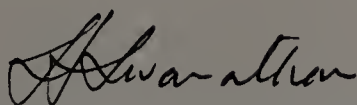
OPTIMAL TEST DESIGNS WITH CONTENT BALANCING AND VARIABLE
TARGET INFORMATION FUNCTIONS AS CONSTRAINTS

A Dissertation Presented

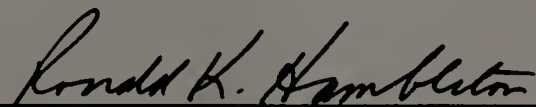
by

LAM TIT LOONG

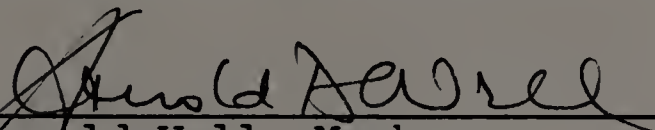
Approved as to style and content by:



Hariharan Swaminathan, Chair



Ronald Hambleton, Member



Arnold Well, Member



Bailey W. Jackson, Dean
School of Education

DEDICATION

This dissertation is dedicated to the Lord Jesus, my Saviour for His wisdom and mercy and for making all things possible for me.

My love goes to my dearest wife, Yeen, and to my lovely children, Evangeline and Zachary who have been supportive and understanding in the course of my studies.

ACKNOWLEDGEMENTS

I am thankful to all who had helped me in one way or other to make my studies possible. My heartfelt thanks go to my committee members, Professor Swaminathan, Professor Ron Hambleton and Professor Arnold Well, whose professionalism have challenged me to greater academic pursuits.

I am also indebted to Professor Sim Wong Kooi of the National Institute of Education, Singapore who had strongly supported my sponsorship for this study, without whom I would not have the financial resources to pursue my academic goals.

Last but not least, my heartfelt thanks go to Peggy Louraine who had painstakingly prepared my thesis.

ABSTRACT

OPTIMAL TEST DESIGNS WITH CONTENT BALANCING AND VARIABLE
TARGET INFORMATION FUNCTIONS AS CONSTRAINTS

FEBRUARY 1993

LAM TIT LOONG, B.SC (HONS.) UNIVERSITY OF LONDON

M.ED, UNIVERSITY OF SINGAPORE

Ed.D., UNIVERSITY OF MASSACHUSETTS

Directed by: Professor Hariharan Swaminathan

Optimal test design involves the application of an item selection heuristic to construct a test to fit the target information function in order that the standard error of the test can be controlled at different regions of the ability continuum. The real data simulation study assessed the efficiency of binary programming in optimal item selection by comparing the degree in which the obtained test information was approximated to different target information functions with a manual heuristic. The effects of imposing a content balancing constraint was studied in conventional, two-stage and adaptive tests designed using the automated procedure.

Results showed that the automated procedure improved upon the manual procedure significantly when a uniform target information function was used. However, when a peaked target information function was used, the improvement over the manual procedure was marginal. Both procedures

were affected by the distribution of the item parameters in the item pool.

The degree in which the examinee empirical scores were recovered was lower when a content balancing constraint was imposed in the conventional test designs. The effect of uneven item parameter distribution in the item pool was shown by the poorer recovery of the empirical scores at the higher regions of the ability continuum. Two-stage tests were shown to limit the effects of content balancing. Content balanced adaptive tests using optimal item selection was shown to be efficient in empirical score recovery, especially in maintaining equiprecision in measurement over a wide ability range despite the imposition of content balancing constraint in the test design.

The study had implications for implementing automated test designs in the school systems supported by hardware and expertise in measurement theory and addresses the issue of content balancing using optimal test designs within an adaptive testing framework.

TABLE OF CONTENTS

	<u>Page</u>
ACKNOWLEDGEMENTS	v
ABSTRACT	vi
LIST OF TABLES	xi
LIST OF FIGURES	xii
Chapter	
1. INTRODUCTION	1
1.1 Statement of the Problem	1
1.2 Purpose of the Study	6
1.3 Theoretical Framework	7
1.3.1 Item Information Matrix as a Basis for Test Design	7
1.3.2 Two-stage and Adaptive Test Designs	9
1.3.3 Use of Binary Programming for Test Design .	10
1.4 Significance of the Study	11
1.5 Scope and Delimitations of the Study	13
2. LITERATURE REVIEW	15
2.1 Classical Test Theory and its Application in Test Designs	15
2.1.1 Classical Test Theory	16
2.1.2 Application of Classical Test Theory in Test Designs	17
2.2 Application of Item Response Theory in Test Designs	20
2.2.1 Item Response Logistic Models	21
2.2.2 Item Pool Assessment Procedures for Test Designs	23
2.2.3 Application of IRT in Test Development	30
2.2.4 Issues Relating to IRT Based Test Designs .	38
2.3 Test Design by Binary Programming	41
2.3.1 Structured Optimal Item Selection	42
2.3.2 Simultaneous Test Construction	43

2.3.3	Minimax Model of Test Construction	44
2.3.4	Maximin Model of Test Construction	46
2.3.5	Development of Two-stage Tests	47
3.	METHOD	49
3.1	Data Source	49
3.2	Item Pool Calibration	49
3.3	Assessing Model-data Fit	50
3.4	Test Development	51
3.4.1	Conventional Tests	52
3.4.1.1	Broad-range Conventional Tests	52
3.4.1.2	Peaked Conventional Tests	55
3.4.1.3	Conventional Tests with Content Balancing	55
3.4.2	Two-stage Tests	57
3.4.3	Adaptive Tests	58
3.5	Scoring	59
3.6	Statistical Analysis	61
3.6.1	Information Analysis	62
3.6.2	Analysis of Score Differences	63
3.6.3	Correlational Analyses	64
4.	RESULTS	65
4.1	Unidimensionality Assessment	65
4.2	Descriptive Statistics	67
4.3	Comparison of OTD and UD Designed Broad-range Tests	68
4.4	Comparison of OTD and UD Designed Peaked Tests ..	69
4.5	Comparison of Content Balanced Conventional Tests.....	71
4.6	Comparison of Tests with Content Balancing Constraint	72
4.7	Comparison of Two-stage Test Designs	76
4.8	Comparison of Adaptive Test Designs	78

5. DISCUSSION AND CONCLUSION	84
5.1 Conventional Test Designs	84
5.2 Two-stage Test Designs	86
5.3 Adaptive Test Designs	87
5.4 Possible Applications of Automated Test Designs in the Schools	88
5.5 Conclusion	89
5.6 Suggestions for Further Study	92
APPENDIX: ITEM BANK PARAMETERS	94
BIBLIOGRAPHY	100

LIST OF TABLES

Table		<u>Page</u>
1.	Fit statistics for linear and nonlinear factor models	67
2.	Analysis of Standardized Residuals for the 1-, 2- and 3-Parameter Logistic Models	68
3.	Distribution and Descriptive Statistics of Test Items by Content in Item Pool	69
4.	Obtained and Target Information Functions of Specified Ability Levels for Conventional Test Designs	73
5.	Correlation of Conventional Test Scores with Standardized Raw Scores	75

LIST OF FIGURES

Figure		<u>Page</u>
1.	Plot of Eigenvalues of Inter-Item Correlation Matrix	66
2.	Obtained Test Information Functions of UD and OTD Designed Broad-range Conventional Tests	70
3.	Obtained Test Information Functions of UD and OTD Designed Peaked Conventional Tests	71
4.	Obtained Test Information Functions of UD and OTD Designed Conventional Tests with Content Balancing	72
5.	Obtained Test Information Functions of OTD Designed Conventional Tests with and without Content Balancing	74
6.	INACC Plots for Conventional Tests with and without Content Balancing	76
7.	RMSD Plots for Conventional Tests with and without Content Balancing	77
8.	INACC Plots for Two-stage Tests with and without Content Balancing	78
9.	RMSD Plots for Two-stage tests with and without Content Balancing	79
10.	Testlet Target Information Bars	80
11.	INACC plots for Adaptive Tests with and without Content Balancing	82
12.	RMSD Plots for adaptive Tests with and without Content Balancing	83

CHAPTER 1

INTRODUCTION

Optimal test design involves the selection of items based on the assumption of the additive property of item information in Item Response Theory from which the standard error of a test can be controlled at different regions of the ability continuum. The choice and the level of difficulty of the items selected by the particular item selection heuristic depends on the anticipated ability distribution of the group of examinees to be tested and the test specification table used. Tests designed for scholarship awards for example, will comprise items of the appropriate difficulty level in which high ability examinees will have a probability of 0.50 of answering the items correctly.

1.1 Statement of the Problem

A common practice among practitioners in designing norm-referenced tests is to select items with difficulties (proportion correct) centered around 0.5 to maximize internal consistency reliability and to maximize test score variance (Allen and Yen, 1979). The test will have most of its items concentrated at one difficulty region and will measure very well, individuals whose ability levels are near this difficulty region of the test. This conventional test is said to be 'peaked' at this particular band (McBride,

1976). Individuals further below and above that level will be measured less precisely by the test. On the other hand, if the test developer should choose items that spread evenly from the lowest to the highest difficulty level, the items will be spread thinly at each difficulty level because of constraints laid by the fixed length of the test.

Consequently, although there is almost equal measurement precision at each ability level, because of the few items located for each ability level, the overall measurement precision is low (McBride, 1976).

However, a more important issue in classical test design is that the item characteristics (item difficulty and item discrimination) depend on the particular examinee samples in which they are obtained (Hambleton and Swaminathan, 1985). Because of this, an item bank calibrated in the classical mode and from which tests are developed is only appropriate if the examinees to be tested are similar in ability distribution to that of the calibration sample.

A better solution to the problem of test construction involves an application of Item Response Theory (IRT) whereby, items from an item pool with known characteristics are optimally selected to fit the target information functions specified for the test. Because of the fact that IRT item parameter estimates are independent of the group of examinees used from the population of examinees from whom the test was designed (Hambleton & Swaminathan, 1985), this

makes the development of an item bank using IRT model more meaningful. Another important feature of IRT is the concept of test information which is inversely related to the standard error associated with the ability estimate (Hambleton and Swaminathan, 1985; page 104). The test information consists entirely of independent and additive contributions from the individual item information. It is this additive property that forms the basis for modern test design. This is in contrast with classical test theory where it is not possible to identify the contribution of an individual item to test reliability or validity independent of the contributions of the other items.

A standard procedure for test design based on the IRT model is described by Birnbaum (1968) which involves setting up a target test information in which the test is to be built and selecting items with item information that will fill the area under the target information. The individual item information are added cumulatively with back-tracking if necessary in order to fill the whole target information curve. Although test designs based on target information is an advantage over that of the classical model, rules for optimal item selection appear to be lacking from literature (Boekkooi-Timminga, 1992). One such contribution on item selection heuristics based on the Birnbaum (1968) procedure is given by Hambleton and Swaminathan (1985). In Birnbaum's (1968) and Lord's (1980) description of the heuristics involved in item selection, it is assumed that the selection

process is done by hand. Hambleton, Arrasmith and Smith (1987) have shown how shorter, yet more efficient criterion-referenced tests can be constructed from a 249-item certification exam based on optimal item selection at the cut-off score of interest.

It can be seen that with a large item pool and with constraints such as the imposition of content balancing, Birnbaum's method of test construction have certain limitations. Firstly, the method involves a manual procedure and it can be time consuming especially when dealing with a large item pool calibrated using the three-parameter logistic model. Secondly, there is no guarantee of optimal results within the constraint of a fixed test length. Thirdly, it is difficult to apply when constraints such as content balancing and administration time are added in the test development process (Boekkooi-Timminga, 1992).

A linear programming approach applied to Birnbaum's method of test design was recently developed and implemented in a number of studies (e.g. van der Linden, 1987, Theunissen, 1985, 1986). Theunissen's (1986) and Adema's (1990) use of binary programming enables the test developer to build a test by first, setting the target information of the test and then proceeding to select items based on specific binary programming algorithms. These studies dealing with a host of item selection algorithms to cater to various test designs have shown that with automated test design, much time is saved and in most cases, optimal

results are achieved. The computer program, Optimal Test Design (OTD) (Verschoor, 1991) was developed for optimal item selection based on the 1- and 2-parameter item response logistic models. The program was subsequently updated to include the 3-parameter logistic model.

A number of factors have to be considered in the test development process. One has to consider the appropriate height of the target information in relation to test lengths. Setting too high a target test information will indeed, ensure a high precision of measurement provided that there are enough good items in the item pool for selection. So, although the development of binary programming procedures allows for fast automated item selection within the computer environment, the whole process is still limited by the characteristics of the item pool. In the use of OTD, the program will register a non-solution problem if there are not enough items from the pool to fit the target test information. Since test designs based on the binary programming approach make use of a set of constraints in the enumeration of a design problem, the success or failure of such a numerical procedure depends ultimately on the distribution and stratification of the item pool. For example, the imposition of content balancing may add further constraints to the test development process if the item characteristic distribution is not homogeneous across content subdomains. Hambleton, Arrasmith and Smith (1987) have shown that content balanced 20-item tests have slightly

lower test information compared to noncontent balanced tests. This is due to the fact that the imposition of additional constraints such as content balancing will mean that the item bank has to be stratified according to the content subdomains of the test. If the distribution of item characteristics such as item difficulty and item discrimination is not homogeneous across content subdomains, poorer quality items may have to be selected across content subdomains to accommodate the content balancing requirement. Although automated test designs have proven to be fast and efficient, comparisons between such techniques with Birnbaum's (1968) manual procedure have yet to be made in order to ascertain the degree in which the resulting test designs approximate to the target information. This study attempts to address these issues.

1.2 Purpose of the Study

The study concerned the development of conventional, two-stage and adaptive tests from an item pool using optimal item selection techniques. Specifically, the main goal of the study was to investigate the influence of variable target test information and content balancing on the outcome of test designs based on the binary programming approach and to examine the measurement precision of these tests. The criterion for ascertaining measurement precision was based on the comparison of obtained test information curves as well as the degree in which the known abilities of the

examinees were recovered by the test designs. Specifically, the goals of this study were:

- 1) To compare the accuracy of automated test designs based on the binary programming approach with Hambleton and Swaminathan's (1985) optimal item selection heuristics in order to ascertain to what extent such procedures approximate closer to the target information.
- 2) To determine and compare the measurement precision of automated test designs with target test information and content balancing as design variables.
- 3) To ascertain whether the imposition of content balancing constraints in the test design process will incur a loss of measurement precision and relative efficiency.

1.3 Theoretical Framework

The following concepts form the bases for the theoretical framework of this study:

1.3.1 Item Information Matrix as a Basis for Test Design

Central to the application of IRT to adaptive testing is the use of item information function as the basis for item selection. According to Birnbaum (1968), for any binary item i , the item score u_i has a Bernoulli distribution. For any fixed value of ability θ , the parameter $P_i(\theta)$ is the probability in which the examinee

gets the answer correct for item i . The item information function (Lord and Novick, 1968) for item i is given as:

$$I(u_i, \theta) = \frac{[P'_i(\theta)]^2}{P_i(\theta) [1 - P_i(\theta)]} \quad (1)$$

The information function is derived from the maximum likelihood function for θ based on the observed item responses, u_i . This function is inversely proportional to the square of the length of the asymptotic confidence interval for estimating ability θ from examinee score y . On the condition that local independence of item responses is kept, the item information is additive such that a test comprising a set of items will have the test information given by the summation of the item information:

$$I(u_1, \dots, u_n, \theta) = \sum \frac{[P'_i(\theta)]^2}{P_i(\theta) [1 - P_i(\theta)]} \quad (2)$$

Using a set of ability values and the corresponding set of items in the item pool, an item information matrix or information table (Thissen & Mislevy, 1990) can be created and stored in the computer. The information table is used for test designs in which a target information for the test is specified and the items are selected. By creating a reasonably large item pool where items are uniformly distributed so that good discriminating items are found in a broad spectrum of difficulty levels, a broad range of ability levels can be measured with good precision.

1.3.2 Two-stage and Adaptive Test Designs

A two-stage testing procedure consists of a routing test whereby, examinees are given a short test which will route them to a second stage measurement test (Lord, 1980). This second stage test consists of a series of peaked tests, each with maximum information at increasing levels of ability. Examinees routed to the appropriate second stage tests will have their abilities estimated more precisely since they are given a test which has maximum information about their ability levels. The number of second stage tests is determined by considerations of economy (Lord, 1980) and by the size of the item pool.

It can be seen that the two-stage test is a simplified version of an adaptive test. The test is adaptive only at one stage - that of routing the examinees to the appropriate second stage measurement test.

In an adaptive test, every individual is administered a different set of test items based on the individual's prior responses. The easier second item is selected from the item pool if the examinee fails the first item and a harder item is selected if the examinee passes the first item. This form of testing differs from the conventional paper-and-pencil tests in which all examinees are administered identical test items. In a sense, adaptive testing is a case of tests designed for each individual examinee (Boekkooi-Timminga, 1992).

1.3.3 Use of Binary Programming for Test Design

Standard IRT test construction practice involves selecting the a number of items from an item pool that will fit the target test information within certain constraints such as content specifications imposed. The above test design problem can be translated into a linear programming problem. A linear programming model formulated to solve a test construction problem attempts to optimally select a number of items in the test subject to the constraints that at least a certain amount of information is obtained at some pre-specified ability levels. This model is stated as follows:

Minimize:

$$\sum_{i=1}^I x_i \quad (3)$$

Subject to:

$$\sum_{i=1}^I I_i(\theta_k) x_i \geq T(\theta_k) \quad (4)$$

so that:

$$x_i \in [0,1] \quad (5)$$

In the above model, x_i is the decision variable for the i th item in the bank where $i = 1, 2, \dots, I$. If $x_i = 1$, the item is included in the test. If $x_i = 0$, the item is not included in the test. $T(\theta_k)$ is the target information value at the ability level θ_k .

The main purpose of this problem is therefore, to load a test with the minimum number of items from a bank so that at a number of θ points, the information $[I_i(\theta_k)]$ in the test is above the target. This general optimization problem can be applied to any test design with constraints imposed including that of two-stage and adaptive tests as can be seen later. Solution of this problem is done by an algorithm called the simplex algorithm, implemented in most computer programs and in particular, OTD (Verschoor, 1991).

1.4 Significance of the Study

Given a calibrated item pool, a test constructor has two general considerations when developing a test. Firstly, he/she has to consider the goal of the test. For example, if the goal is to select gifted candidates for scholarships, then only a certain percentage of the difficult items in the pool is selected in order that the most gifted has a 0.5 probability of getting the items correct. Using Birnbaum's (1968) method, the test constructor will set a higher target information function at the appropriate criterion region of the ability continuum. Secondly, the choice of the items is constrained by the test length as well as the test specifications such as content emphasis and item format. The process of optimal item selection can be done manually although it is time consuming and might only yield an approximate solution to the test design problem after several back-tracking cycles (van der Linden & Boekkooi-Timminga, 1989). This study attempts to compare the

accuracy of automated test designs with Birnbaum's (1968) and Lord's (1980) manual procedure of enumerating a test design problem.

The test constructor with a knowledge of the concept of IRT test design need not know the intricacies of Optimization Theory since the application of the theory is translated into computer codes. This study highlights the relative ease and speed in which different kinds of tests can be developed from the same item pool.

In the school setting, the teacher in implementing an instructional program normally has specific goals and skill areas in mind. Content balancing is important to school testing programs where the test specification table plays an important role in delineating the subject matter to be tested. Where there is a need to make use of IRT in the school setting, the use of a properly designed test will satisfy the requirements of the school testing program.

The success of automated item selection depends ultimately on the quality of the item pool. Maintenance of such a pool is outside the scope of the computer environment as this relies on the expertise of the subject matter specialist and the skills of the item constructor.

In this regard, this study also attempts to highlight the importance of the item pool characteristics and the proper maintenance and stratification of the pool which will ultimately affect the solution of the test design problem.

The study also points to the importance of the role of the test constructor and subject matter specialist in developing and maintaining an item bank appropriate for the test design. As highlighted by Wainer and Kiely (1987), test designs need a certain measure of 'control' in order that some measure of congruence between the goals of the testing program and the goals of instruction be met.

1.5 Scope and Delimitations of the Study

The study takes the form of a real-data simulation using the item responses of examinees based on the administration of a credentialling exam. The examination paper consists of 250 items which is sufficient to form an item pool for this study. Three limitations are apparent in this regard:

1. The item pool is derived directly from a single administration of an exam paper. The items forming the exam paper were assumed to be appropriately selected from a larger item pool.

As such, the quality of this item pool will depend on the quality of the items in the examination paper.

2. Although the abilities of the examinees are known, their true abilities are unknown.

Recovery of abilities by the tests will be based on the known abilities which have error components of their own. That is, the known abilities are not error-free and any comparison

of measurement precision of the tests is only relative in this sense.

3. As in most simulation studies involving a live dataset, it is assumed that the way in which the examinee responds when the test is presented in different modes is similar.

CHAPTER 2

LITERATURE REVIEW

This chapter is organized into three sections. The first section deals with the development of the Classical Test Theory and its applications in test designs where item selection strategies and their limitations are discussed. The second section deals with the development of Item Response Theory and focuses on how it addresses certain limitations posed by the Classical Test Theory. This section then continues on with the application of Item Response Theory in test designs followed by a discussion on certain issues relating to its implementation. The third section reviews recent applications of binary programming techniques which attempt to complement the application of IRT in the development of conventional, two-stage and adaptive tests, thus forming the background of this research study.

2.1 Classical Test Theory and its Application in Test Designs

Classical test theory was based in part on the early statistical foundations laid by Karl Pearson (1857 - 1936) who developed a number of statistical techniques which formed the core of basic measurement theory (Allen & Yen, 1979; page 3). These include the Pearson product moment correlation coefficient and the chi-square goodness of fit test. The first standardized achievement test was developed

by Binet and Simon in 1905. Work by Charles Spearman (1863 - 1945) led to the modern concepts of test reliability and factor analysis.

2.1.1 Classical Test Theory

Classical test theory postulates that an examinee has a true score (T) defined over a domain of test content. This true score is fixed but if the person is tested more than once, the observed score (X) varies because of variation due to measurement errors. The error scores over examinees are random with mean = 0 and uncorrelated with the true scores. It is assumed that repeated test administrations are independent of each other so that each test has no influence on subsequent tests. Since in reality, this is not possible, T is defined as an "expected" test score and is treated as a theoretical construct. The observed, true and error scores are linearly related. From this definition of classical test theory, the following is a model and a set of assumptions (Allen & Yen, 1979):

Model : $X = T + E$

Assumption 1: $E_{\text{mean}} = 0$

The error scores over examinees on a single test administration is zero.

Assumption 2: $\rho_{ET} = 0$

The error scores and the true scores obtained by a population of examinees on a test administration are uncorrelated. This implies that examinees with high true scores do not have systematically more positive or

negative errors of measurement than examinees with low true scores.

Assumption 3: $\rho_{E_1, E_2} = 0$

The error scores for two different tests are uncorrelated. That is, if a person has a positive error score for Test 1, he/she is not more likely to have a positive or negative error score on Test 2.

Assumption 4: $\rho_{E_1, T_2} = 0$

This assumption states that the error scores on one test are uncorrelated with the true scores on another test.

It can be seen from the above assumptions that the error of measurement in the classical sense, is an unsystematic, or random deviation of an examinee's observed score from a theoretically expected observed score.

Two tests (denoted by "1" and "2" below) are said to be "parallel" if:

- a) they measure the same content,
- b) $T_1 = T_2$ for each examinee and
- c) $\sigma^2(E_1) = \sigma^2(E_2)$ (error variances on the two tests are equal).

In its simple form, the reliability of a test is the correlation of the observed scores ($\rho_{xx'}$) on a parallel test.

2.1.2 Application of Classical Test Theory in Test Designs

Classical test designs are based on two central concepts - test reliability and test validity.

Based on the assumptions of Classical Test Theory, the concept of test reliability can be further derived (Lord, 1980; pages 4 & 5). The relationship between reliability, error score variance and observed score variance is given by:

$$\rho_{xx'} = 1 - \sigma_E^2 / \sigma_x^2 \quad (6)$$

It is from Equation 6 that the quantity, coefficient alpha (α) is obtained (Gulliksen, 1950):

$$\rho_{XT}^2 = \rho_{xx'} \geq \frac{n}{n-1} \left(1 - \frac{\sum \sigma_i^2}{(\sum \sigma_i \rho_{ix})^2} \right) \quad (7)$$

α is the lower bound of the reliability coefficient. σ_i^2 is the item variance and ρ_{ix} is the item-test correlation (or item discrimination).

For binary items, the item variance can be obtained from the item difficulty, p_i (or proportion correct) and is computed as $p_i(1-p_i)$.

Test validity is defined as:

$$r_{xy} = \frac{\sum \sigma_i \rho_{iy}}{\sum \sigma_i \rho_{ix}} \quad (8)$$

where ρ_{iy} is the item-criterion correlation.

Given a pool of test items the test developer who wants to design a test that has maximum reliability will:

1. select items with large item-test correlations
(in order to maximize the denominator of Equation 7 so that α is increased) and
2. increase the test length.

On the other hand, a test developer who wants to design a test that has maximum validity will:

1. select items with large item-criterion correlations and low item-test correlations and
2. increase the length of the test or the criterion used.

This poses a dilemma for the test constructor who wants to maximize both the validity and reliability of the test because both large and small discriminating items will then be desirable. The test developer will then have to decide which goal is more important in order to determine the method of item selection bearing in mind that the test built on an emphasis of either goals will have different combinations of items. That is, if the items are chosen to maximize validity, the resulting test will not have good reliability.

Hambleton and Swaminathan (1985; pages 1 - 3) listed a number of shortcomings in the Classical Test Theory which are fundamental to measurement and test designs. Among these are:

1. Both reliability and validity indices used in the classical model are group dependent and therefore have limited generalizability. This is because the item difficulty and item discrimination used in both indices depend on the particular examinee samples in which they are obtained. The item discrimination index will increase when obtained

from a more heterogeneous sample. Hence, item statistics are useful only in item selection when constructing tests for examinee populations that are very similar to the examinee sample in which the statistics were obtained.

2. The concept of test reliability is defined in terms of parallel forms which is difficult to apply in practice since a number of factors come into play when individuals are administered a test the second time.
3. Standard errors used in the classical sense are averaged standard errors which are averaged over the ability levels so that every examinee is presumed to have the same error variance which might not be true in a testing situation where individual differences such as consistency and moods interact with ability levels when performing tasks.

2.2 Application of Item Response Theory in Test Designs

The solutions to the problems highlighted in the previous section come in the application of Item Response Theory (IRT). The use of IRT makes it possible to estimate trait levels from the responses to a series of items (Weiss, 1982). Credit is given to Lord's (1970) work in laying the psychometric foundation for applying IRT concepts to test designs.

2.2.1 Item Response Logistic Models

Birnbaum's (1968) three-parameter logistic model assumed that the latent trait, θ is unidimensional with an unrestricted domain, $-\alpha < \theta < \alpha$. It is also assumed that the principle of local independence holds (Lord and Novick, 1968) where for a fixed value of θ , the distributions of the item scores are independent of one another.

For item i , and the corresponding item response, u_i , the conditional distribution given θ of a single item response is $L(u_i|\theta) = P_i(\theta)$ if $u_i = 1$ and $L(u_i|\theta) = Q_i(\theta)$ if $u_i = 0$. The response vector, $\mathbf{v}' = (u_1, u_2, \dots, u_n)$ where u_i is scored either 1 or 0 is such that the likelihood function for estimating an individual's latent trait based on this response pattern is:

$$\Pr(\mathbf{v}|\theta; \mathbf{a}, \mathbf{b}, \mathbf{c}) = \prod_{i=1}^n P_i(\theta)^{u_i} Q_i(\theta)^{1-u_i} \quad (9)$$

where:

$$Q_i(\theta) = 1 - P_i(\theta) \quad (10)$$

Equation 9 is viewed as the conditional distribution of the pattern \mathbf{u} of item responses for a given individual with ability θ and for known item parameters, \mathbf{a} , \mathbf{b} , \mathbf{c} . The u_i are random variables and since they can be determined from the examinees' answer sheets, they become known constants. θ , \mathbf{a} , \mathbf{b} , and \mathbf{c} are considered fixed. If the item parameters are known from pretesting, Equation 9 becomes a function of the mathematical variable, θ and is considered as the

likelihood function for θ . The maximum likelihood estimate of the examinee's ability is the value of θ that maximizes the likelihood of his/her observed responses u_i .

The item characteristic function for the three-parameter model is then represented by:

$$P_i(\theta) = c_i + \frac{1 - c_i}{1 + e^{-Da_i(\theta - b_i)}} \quad (11)$$

where: D is a scaling factor given the value of 1.7,
 a_i is the item discrimination,
 b_i is the item difficulty and
 c_i is the pseudoguessing parameter.

The parameter c_i is the lower asymptote of the item characteristic curve and represents the probability of the examinee with low ability correctly answering the item. If the pseudoguessing parameter is assumed to be zero, then the two-parameter logistic model results. This assumption is most probable if the test is not too difficult, as in the case of competency testing following effective instruction (Hambleton & Swaminathan, 1985). The one-parameter (or Rasch) model results if all items have equal discriminating power and guessing is assumed to be zero.

Where there is a close fit between the item response model and the test data of interest, a number of features in IRT can be seen to be particularly advantageous over the use of Classical Test Theory in test designs (Hambleton and Swaminathan, 1985; pages 10 & 11):

1. It places the person trait levels on the same scale as the item difficulty so that item selection can be appropriately done by matching the ability estimate with the difficulty of the item. The classical item difficulty value (p) is not just a function of the difficulty of the item alone, but a function of the examinee characteristics as well.
2. Examinee ability estimates are independent of the choice of test items used from the population of items which were calibrated. That is, the items are treated as fungible (interchangeable) units and that responses to the items are independent of each other so that ability levels can be estimated based on subsets of items administered to the individual. This enables the development of tests with items selected from a calibrated item pool.
3. Items can be selected not just on difficulty levels alone, but on discrimination and pseudoguessing (as in the case of using the three-parameter logistic model) thus, adding more information to the item selection process.

2.2.2 Item Pool Assessment Procedures for Test Designs

Since IRT assumes unidimensionality to account for examinee performance in a single trait, evidence of unidimensionality must be ascertained in an item pool from

which a test is built. Unidimensionality depends a lot on the item selection process (Green and associates, 1984). Urry's (1981) suggestion for selecting items of at least 0.80 discrimination means in psychometric terms that the items will have a higher correlation with the underlying trait they are measuring; thus ensuring unidimensionality. This is true of both conventional or adaptive tests. However, although selection of items with high α -values also means providing for greater information, Green and associates (1984) commented that this might mean rejection of some good item types as well as items that measure some important content areas. A more compromising criterion for accepting unidimensionality as suggested by Green and associates (1984) is to accept the factor pattern where there is one prominent factor that accounts for 70% of the total common variance even though there may be secondary factors.

However, there are a number of fundamental problems associated with the classical linear factor analysis solution. Firstly, linear factor analysis assumes that the relationship between the observed variables and the underlying factors is linear and that the variables are continuous in nature. In the majority of binary item responses, the relationship between the item responses and the underlying trait is nonlinear and that these observed variables are categorical. The application of linear factor analysis to binary responses results in an approximation of

the nonlinear relationship to a linear one. One result is that difficulty factors emerge if guessing is allowed (Hulin, Drasgow & Parsons, 1983).

In an attempt to solve this problem, McDonald (1980, 1982) demonstrated that applying nonlinear factor analysis to unidimensional binary data results in nonlinear factors instead of difficulty factors. Since the latent trait is related to performance in a nonlinear fashion, the application of nonlinear factor analysis seems appropriate. McDonald's procedure is implemented in the computer program, NOHARMII (Fraser, 1983).

Another approach to solving the problems associated with linear factor analysis is to make use of the full information approach to item factor analysis. This method avoids the use of interitem correlations since the classical factor analytic model is not suitable for binary variables such as the item score (Mislevy, 1986). Factor loadings are estimated directly from the response data beginning with one factor and increments in goodness-of-fit of the model are tested for additional factors entered in the model. The analysis continues until the addition of factors is not met with a significant increase in goodness-of-fit. A computer program, TESTFACT (Wilson, Wood, & Gibbons, 1984) is designed to handle this analysis.

Green and associates (1984) suggest that a simpler way for analysis of an item pool in which the items are clustered in different content areas, is to score each

subtest and correlate the subtest scores. If the corrected correlations (i.e. corrected for disattenuation due to unreliability) are about 0.9 or higher, then unidimensionality of the item pool can be accepted. This is applicable in the case of large items pools of say, a few hundred items where items are categorized into a number of content areas. Each content area will therefore represent a subtest.

A nonparametric procedure for assessing dimensionality was developed by Stout (1987). Stout's (1987) based his procedure on the premise that any subpopulation of examinees with approximately equal test scores on a reasonably long test should have equal abilities and thus local independence should be adhered to. On the other hand, if a test is multidimensional, then the examinees with approximately equal test scores may differ widely in the components that form their ability vectors. Stout's method has been shown to be discriminating well between unidimensional and two-dimensional tests in simulated datasets for correlations between abilities as high as 0.70 (Nandakumar, 1991). Previous factor analytic procedures are not appropriate for analyses of large item pools because of limitations on the matrix sizes and heavy computation memory involved. NOHARMII for example, can take in a maximum of only 140 items. Hence, Stout's approach appears elegant for large item pools of 200 items or more since the procedure mainly involves computations of variance estimates of subgroups to

come up with an index for testing the null hypothesis for unidimensionality.

If the item bank is kept without modification for a period of time, effects such as curricular and technological change over time may affect the item bank scale. Such an effect on the item bank parameters is called item parameter drift. This is defined as the differential change in item parameter estimates over time (Goldstein, 1983). For example, Bock, Muraki and Pfeiffenberger (1988) found from the results of a two-way ANOVA (items X year-groups), indications of item parameter drift in Physics Achievement Test (College Board) data. They attributed this to the change in Physics curricula over the 10-year period in which the test was administered. As part of the maintenance process of the item bank, certain items need to be retired when they are deemed to be overexposed and the size of the item bank need to be increased over time by preequating the tryout items to the bank scale. Item parameter drift is possible and a reason advanced by Sykes and Fitzpatrick (1992) is the possibility of declining examinee ability levels over the years with the result that the equating method used does not fully capture this trend.

Other possible reasons for item parameter drift are item position, context effects and item content of tryout items selected to ultimately link up with the item pool scale. Wainer and Kiely (1987) and Whitely and Dawis (1976) have found that item difficulty estimates can vary as a

function of item position. When a prequating procedure is used, the placement of the tryout items into the item bank may affect their item calibrations due to item order effects. In an analysis of a professional licensure exam using ANCOVA methodology, Sykes and Fitzpatrick (1992) found an increase in item pool b values for one of the content categories after controlling for elapsed time between test administrations. If item parameters are influenced by other items in the test, then context effects are occurring. This again, have implications for the calibration of items for item pools. Yen (1980) in her study of seven test forms of the California Achievement Test (1977) found that item parameters estimated from the same context were more highly related than item parameters estimated from different contexts.

Changes in item parameter values due to various factors associated with the item bank maintenance process tend to produce essentially linear transformations of trait estimates (Yen, 1980). These transformations affect the means and standard deviations of the examinee trait values as well as the relative sizes of individual trait values. Bock, Muraki and Pfeiffenberger (1988) proposed a method for maintaining and updating an IRT scale over a period of time while accounting for item parameter drift. This procedure can be extended to maintaining an item pool scale. The procedure which is implemented in the program, BIMAIN (1987) is an extension of the BILOG program. This procedure

involves the estimation of the likelihoods used in estimating the estimated numbers of correct responses and numbers of respondents at the quadrature points (the E step of the EM algorithm) by first excluding the tryout items. After the likelihoods are estimated, these are used to estimate item parameters of the block of items in the item pool together with the tryout items.

Any test design depends on the quality of the calibrated item pool. An item pool of credible size cannot be build using a one time administration of a few hundred items to a single sample of examinees for obvious reasons. Apart from size, a good item pool requires good quality items over a wide ability range. In addition, the assumptions of the psychometric model used in the testing program must be satisfied. Although item calibration using IRT means that item parameters are invariant across population, Green and associates (1984) suggested that the population used for item calibration should be comparable to the target population especially in range. A simple item calibration scheme which made use of a randomized block design for administering 250 items was given by Wainer and Mislevy (1990). This involves dividing the 250 items into 10 sets of 25 items each and administering 10 forms of the test randomly; each form consisting of a non-overlapping set and an overlapping set.

The item response model chosen for item calibration has to be assessed for model-data fit. A number of approaches had been discussed by Hambleton and Swaminathan (1985).

Among these are:

1. Residual analysis of model-fit data in which residual (difference between the observed data and an estimated item characteristic curve) plots across ability groups are made. Fit of the model to the data is judged by the smallness of the residuals or the closeness in which the observed average item performance of each ability group is to the estimated item characteristic curve.
2. Plots of true and estimated item and ability parameters (Hambleton and Cook, 1983).
3. Comparison of observed and predicted score distributions (Hambleton and Cook, 1983).

2.2.3 Application of IRT in Test Development

IRT offers a more meaningful method of item selection over that of the Classical Test Theory for two reasons. Firstly, the item parameters are sample invariant while the success of test designs using the classical method depends on how closely the calibration sample matches the population in which the test is intended. Secondly, the standard error of measurement used in the classical sense is an average error estimate applied to the whole group in which the test was administered, implying that the size of the error of measurement is independent of the 'true scores' of the

examinees taking the test (Hambleton & Swaminathan, 1985; page 236). The IRT counterpart of the classical standard error of measurement is the test information function and its advantage is seen by the fact that the item information function has an additive property, each contributing independently to the test information function. This has important implications for test designs when the target information functions are specified and items are selected independently to fit the area under the information curve. This is not possible with the classical procedure because the contribution of an item to the test reliability cannot be determined independently from all the other items in the test. The test information function accounts for the estimate of the error of measurement (SEM) where $SEM = 1/\sqrt{\text{Information}}$ for each ability level instead of giving each examinee the same group error estimate in the classical sense.

Lord (1980) outlined Birnbaum's (1968) procedure for the use of item information functions in test designs as follows:

1. Describe the shape of the target information function in which the test is to be built.
2. Select the items with item information that will fill up the hard-to-fill areas under the target information function.

3. After each item is added to the test, calculate the test information function for the selected test items.
4. Continue selecting the items until the test information function approximates the target information function satisfactorily.

Hambleton and Swaminathan (1985) illustrated the application of Birnbaum's procedure in test design by making use of a hypothetical example of a pool of 12 items. After specifying a target information of 6.25 from -2.00 to +2.00 on the ability scale, items which supply a larger amount of information over a broad ability range was first chosen and the obtained test information plotted. Items with high information over a narrower ability range were then selected to fill the hard-to-fill areas under the target information curve. In another study, Hambleton and Swaminathan (1985) compared the efficiency of five item selection procedures in the construction of a scholarship selection test and a grading test:

1. Random: A table of random numbers were generated and items were selected based on the random numbers.
2. Standard: Items were chosen based on classical p-statistic between .30 and .70.
3. Low/Middle/High difficulty: The best items with maximum information at the ability level of interest were chosen.

4. Up-and-down: An item with the highest information at the lowest specific ability level of the target information function was chosen. The items with the highest information were chosen from each of these specified ability levels upwards and the cycle repeated until the target information levels were reached.

In the development of a scholarship test where the target information was set at the high end of the ability continuum, the authors found that the up-and-down method provided maximum information over a broader range of abilities. The random and standard methods were found to be inferior. In the development of a grading test where the target information was bimodal, the low-high difficulty method was found to be most appropriate.

A two-stage testing procedure consists of a conventional routing test followed by a number of conventional second-stage measurement tests. The administration of the second-stage test depends on the examinee's score on the routing test. The main advantage of such a testing procedure is that the difficulty level of the second test is matched to the ability level of the examinee (assuming that the routing test performs its function well). As such, the test adapts only once - that at the second stage.

Lord (1980) investigated over 300 two-stage test designs of different test lengths using a heuristic applied

to Birnbaum's (1968) procedure. Among some of his conclusions were:

1. If the routing test is too long, not enough items are left for the second-stage test. As such, the routing test functions best as a single conventional test rather than having to rout the examinees to the second-stage level which have poorer measurement precision. On the other hand, if the routing test is too short, then examinees are likely to be poorly allocated to the second-stage measurement tests.
2. At least four second-stage tests covering the range of the ability spectrum were needed for effective measurement.

Lord (1980) without the benefit of computing power used arbitrary and fixed item difficulties as part of his item selection heuristic. For example, in the 60-item two-stage test designs, he designed four second-stage tests, each with the same difficulties, $b \pm 1.00/a$ and $b \pm 0.50/a$ where a is a fixed value. From Lord's (1980) study, two-stage tests were shown to be efficient in measuring examinees at the extremes of the ability range although they may not be as effective as the adaptive test in measuring the same ability regions. Again, without the benefits of automated testing within the computer environment, Lord (1980) suggested various ways of administering the routing test such as self-scoring of the paper-and-pencil test and the immediate

administration of the appropriate second-stage test by the test administrator after knowing the routing test score. Lord (1980) was apparently not too concerned about the honesty of the examinee in self-scoring and suggested that the effect of a routing test scored improperly was 'simply to lower the accuracy of the final second-stage score of the examinee' (Lord, 1980; page 140).

Modern IRT-based adaptive testing involves an item pool from which items are selected in the test administration process. The pool generally consists of highly discriminating items, equally distributed across trait levels. The items are calibrated for difficulty, discrimination and guessing (Lord & Novick, 1968). A requirement for IRT analysis of the item pool is that the item responses are locally independent and this is tested by ascertaining unidimensionality of the items. Urry (1977) suggested that an item bank designed for CAT must have the following requirements:

1. item discrimination must exceed 0.8
2. item difficulty must have a rectangular distribution from -2.0 to +2.0
3. item parameters for guessing should be less than 0.3 and
4. item pool must have at least 100 items.

Weiss (1985) suggested an item pool of 150 to 200 items for optimum results in CAT. However, CAT had been adapted from conventional tests by just using the items from the

fixed length tests. This was done by selecting items using maximum information strategy until there is no items left at the current trait level (Weiss, 1982; Weiss & Kingsbury, 1984).

With the advent of high power, but relatively cheap desktop computers such as the 32-bit "486" machines with video graphics capabilities, CAT is enhanced with the possibility of a wide range of perceptual and visual tests. However, no computer system will enhance CAT without the necessary software. According to Weiss (1985), a typical CAT software must be able to create and update an item pool, create instructional sequences to make the adaptive test user friendly, select items by IRT procedures, terminate the test based on the particular strategy used, estimate individual trait levels, store test data, and produce test interpretations and test reports.

Generally, item selection strategy involves the following:

1. The initial estimate of the examinee's ability level is obtained. In many instances, the estimate of 0.0 is given.
2. This initial ability estimate is used to select an appropriate item from the item pool.
3. From the response of the examinee, the item is scored and the item score is used to revise the estimate of the examinee's ability level.

4. From the new estimate of the examinee's ability, the next item, appropriate to the examinee's new trait level is selected from the item pool and the process is repeated.
5. Based on an acceptable precision of the trait level estimate or unavailability of items pertinent to that trait level, the test is terminated.

An item is then selected from the pool that has the maximum information possible to measure that particular trait level. After the administration of the item, the new trait level estimated from the response to that item is used to select another item in the pool, whose information function is most appropriate for the new trait level. Two common procedures for scoring response vectors in adaptive testing are maximum likelihood and Bayes modal estimation (Wainer & Mislevy, 1990).

Two common criteria are used for termination of the CAT procedure. The first involves a preset standard error of estimate (SEE). This arbitrarily selected value will yield some expected level of validity given by:

$$\rho_{\theta\theta} = \sqrt{(1 - \sigma_m^2)} \quad (12)$$

The second criterion involves a specific number of items that have been administered and termination is done regardless of σ_m^2 . One problem associated with maximum likelihood scoring is that ability estimates cannot be

determined for response patterns in which the examinee answers all the items correctly or all the items incorrectly. There are also some unusual kinds of response patterns in which the maximum likelihood estimation procedure fails to converge.

Samejima (1973) argued that there is no unique solution of θ which satisfies every possible response pattern. That is, the maximum likelihood function does not always provide a unique maximum likelihood estimate unless a subdomain of the latent trait such that $\max(\theta_i) < \theta < \alpha$ is considered such that $\max(\theta_i)$ is the maximum value of θ_i for $g = 1, 2, \dots, n$, and the left hand part of the ability domain is left out. However, Lord (1980) noted that this problem did not usually arise when large item pools ($n > 20$) were used. The problem of non-unique solutions due to all correct or all incorrect answers is usually solved by utilizing a prior ability distribution as in Bayes modal estimation.

2.2.4 Issues Relating to IRT Based Test Designs

Birnbaum's (1968) description of test design based on the additive property of the item information and the optimal item selection within a target information assumed that the selection is done by hand. That is, although the computer can be used to compute an information matrix of all the items in the pool for the number of θ points specified by the target information, the optimal selection of the items is based on the judgement of the test constructor. In

mastery testing where the target information is high at only one θ level, item selection for minimum test length is relatively straightforward (Hambleton & de Gruijter, 1983). The item information can be sorted from high to low at that particular θ level and the most informative items at that level meeting the target information are selected. However, with conventional tests, as well as two-stage tests (which comprises actually a routing conventional test and a set of second-stage conventional tests), selecting the shortest test to meet the target information over a range of θ levels may not be so easy if done by hand (de Gruijter, 1990). This is especially true if the item pool is large and stratified by content subdomains, item formats and other variables involved in the decision making process. Even if the item pool is not stratified, the manual procedure of item selection can be time consuming, involving a number of backtracking cycles till an optimal solution is reached in order to achieve a test of a specified or minimum length desired by the test constructor. There is also no guarantee of optimal results within the constraint of a fixed test length. It is also quite difficult to apply when constraints such as content balancing and administration time are added in the test development process (Boekkooi-Timminga, 1992).

In the case of adaptive testing by maximum information item selection, every examinee technically takes an individually designed test. Because the computer cannot

read test items which are pertinent to a particular area in the curriculum since they are selected on the basis of item statistics, this gives rise to an imbalance in test content. In a conventional test, this will not arise as the test developer would have used a test specification table to serve as a blueprint for test development and to ensure a balance in the content.

As a result of this, content validity is put into question since the items administered may not follow the test specification table (Wise & Plake, 1989). To solve this problem, the item selection strategy can be modified to take test specification into account apart from examinee ability estimates. Kingsbury (1990) had shown how this could be done using the MicroCAT (Assessment Systems Corporation, 1987) software whereby a pre-selection strategy can be adopted in the software to ensure content balancing.

In an attempt to address the issue of content balancing, Wainer and Kiely (1987) proposed the testlet model in which the examinee is given a fixed number of predetermined paths in a pyramidal item selection procedure. Kingsbury and Zara (1989) however, criticized the use of the testlet model on the grounds that this will reduce measurement accuracy because of the "weak" prestructured selection strategy. Furthermore, the use of pyramidal selection strategy is rather inefficient as it requires a rather large item pool. Kingsbury and Zara (1989) proposed a constrained version of CAT whereby a number of components

are built into the selection algorithm. These include a content balancing algorithm whereby, items selectively take into consideration, the test specification of the test developer. The rationale for the administration of hierarchical testlets instead of single items in CAT is that it has the advantage of limiting context effects and item exposure.

Attempts had been made to apply binary programming in test designs in order to optimize the selection of items appropriate to the ability level of the examinees and to fit the target information curve within the kind of constraints imposed. This is an alternative to the trial and error procedure of Birnbaum (1968). The last section of the review addresses this procedure.

2.3 Test Design by Binary Programming

Yen (1983) originally suggested the use of linear programming techniques for test construction. Although she proposed to optimize an overall-quality index which is a function of item discrimination, fit and bias no explicit optimization model was given. Theunissen (1985) was the first to formulate a binary (0-1) linear programming model for solving test construction problems. Just as in Birnbaum's (1968) procedure, a target information function over a number of θ levels is used. A linear programming model formulated to solve a test construction problem attempts to minimize the number of items in the test subject to the constraints that at least a certain amount of

information is obtained at some pre-specified ability levels. This model is stated as follows:

Minimize:

$$\sum_{i=1}^I x_i \quad (13)$$

Subject to:

$$\sum_{i=1}^I I_i(\theta_k) x_i \geq T(\theta_k) \quad (14)$$

so that:

$$x_i \in [0, 1] \quad (15)$$

In the above model, x_i is the decision variable for the i th item in the bank where $i = 1, 2, \dots, I$. If $x_i = 1$, the item is included in the test. If $x_i = 0$, the item is not included in the test. $T(\theta_k)$ is the target information value at the ability level θ_k . All items are assumed to fit the one-dimensional item response model.

The main purpose of this problem is therefore, to load a test with the minimum number of items from a bank so that at all the specified target θ points, the information in the test is above the levels $[I_i(\theta_k)]$ considered.

Following the above general model, a number of alternative objective functions and constraints have been developed.

2.3.1 Structured Optimal Item Selection

Theunissen (1986) considered the case where it is necessary to construct a test in which the items have to be

sampled from a number of content subdomains. If the number of items in each subdomain is fixed, then for the case of three subdomains involved, the optimal test design above is altered by changing Equation 13 to:

$$\sum_{i=1}^r x_i = n_1, \quad \sum_{i=r+1}^s x_i = n_2, \quad \sum_{i=s+1}^t x_i = n_3 \quad (16)$$

where $n_1 + n_2 + n_3 = N$ (the number of items forming the test). From Equation 16, the number of items in the three content subdomains forming the item bank are r , $(s - r)$ and $(t - s)$ in that order. t is the total number of items in the bank. It is also assumed that the items are originally grouped into the three content subdomains specified above. If the number of items to be drawn from the subdomains is not a fixed constant then proportional drawing of the items can be done by altering equation 16 to:

$$a \sum_{i=1}^r x_i = b \sum_{i=r+1}^s x_i \quad (17)$$

where the ratio of a to b indicates the proportionate item sampling.

2.3.2 Simultaneous Test Construction

This is an extension of Theunissen's (1985) model where T number of tests are constructed at the same time instead of a single test (Boekkooi-Timminga, 1987). Simultaneous test construction is important where parallel tests are needed (van der Linden & Boekkooi-Timminga, 1988). Test are considered to be parallel if their information functions are

the same (Samejima, 1977). The modification on Theunissen's (1985) basic model is as follows:

Minimize:

$$\sum_{i=1}^I \sum_{t=1}^T x_{it} \quad (18)$$

Subject to:

$$\sum_{i=1}^I I_i(\theta_k) x_i \geq T_t(\theta_k) \quad (19)$$

so that:

$$x_{it} \in [0, 1] \quad (20)$$

2.3.3 Minimax Model of Test Construction

Theunissen's (1985) binary programming model for test construction faces a limitation in that the obtained information functions usually have a peak in the middle of the ability interval (van der Linden, 1987). This is because the algorithm will select items with the bulk of their information in the interval $[\theta_1, \theta_k]$ specified by the model. For the 1-P and 2-P logistic model, the item information are symmetric about their difficulty parameter values and the tendency of the algorithm is to select items located in the middle of the interval (van der Linden, 1987). For the case of the 3-P model, because of the presence of the pseudoguessing parameter, the distribution of the item information functions is skewed to the left and

the items somewhat to the left of the interval will be selected.

The minimax model proposed by van der Linden (1987) specifies the minimization of the largest deviation from the target test information subject to condition that all deviations are non-negative. The model is as follows:

Minimize y .

Subject to:

$$\sum_{i=1}^I I_i(\theta_k) x_i - y \leq I(\theta_k) \quad k = 1, \dots, K \quad (21)$$

$$\sum_{i=1}^I I_i(\theta_k) x_i \geq I(\theta_k) \quad k = 1, \dots, K \quad (22)$$

$$x_i \in [0, 1] \quad i = 1, \dots, I \quad (23)$$

y denotes an arbitrary upper bound and $I_i(\theta_k)$ is the value of the information function of item i at the point θ_k . The arbitrary variable y is a dummy variable and does not contain any item or test parameters.

The model specifies that the deviation of the obtained information function from the target information function should not be larger than the upper bound y . The constraint in Equation 21 stipulates that these deviations are non-negative. By minimizing the upper bound y the obtained test information function will tend to conform to the target information, resulting in the smallest possible peak.

2.3.4 Maximin Model of Test Construction

This is an alternative model to van der Linden's (1987) minimax model. The model conceptualized by van der Linden and Boekkooi-Timminga (1989) has the additional potential of controlling test length. The model is as follows:

Maximize y

Subject to:

$$\sum_{i=1}^I I_i(\theta_k) x_i - r_k y \geq 0 \quad k = 1, \dots, K \quad (24)$$

$$\sum_{i=1}^I x_i = n \quad (25)$$

$$x_i \in [0, 1] \quad (26)$$

y now is the lower bound which has to be maximized.

Equation 25 sets the test length to n .

A number of constraints can be added to the maximin model. If test constructor wants to control for test composition, the constraint in Equation 25 can be modified by letting v_j ($j = 1, \dots, J$) be a subset of items in the banks pertaining to a set of content subdomains. The modified constraint is as follows:

$$\sum_{i=1}^I x_i = n_j \quad (27)$$

If the test constructor wants to control for administration time, the length of the test can then be

controlled by specifying the selection of items based on item administration time, t_i . The constraint in Equation 25 is then replaced by:

$$\sum_{i=1}^I t_i x_i \leq T \quad (28)$$

where T is the time limit for administration of the whole test.

2.3.5 Development of Two-stage Tests

Two-stage testing previously defined can be developed by the application of either the Theunissen (1985) minimization model or the maximin model of van der Linden and Boekkooi-Timminga (1989).

Theunissen (1985) suggested the use of a small subset of items from an item pool to be used as the routing test. This selection can be done by the specification of a target information function and the application of the minimization model.

An additional constraint can be added to Theunissen's minimization model:

$$\sum_{i=1}^n x_i = n \quad (29)$$

where n is the number of items to be selected.

In the development of the second-stage test, a number of sequential segments specified by the θ levels on the ability continuum are selected based on the desired number of second-stage tests. For a fixed number of items in each

second-stage test, the same set of equations for the minimization model is used to solve the problem.

Adema's (1990) procedure for development of the routing test is the same as the development of any conventional test using the maximin model. In the development of the second-stage tests, the maximin model is modified by giving an additional constraint as follows:

Maximize y

Subject to:

$$\sum_{i=1}^I I_i(\theta_k) x_i - r_k y \geq 0 \quad (30)$$

$$\sum_{i \in U} x_i = 0 \quad (31)$$

where U is the set of items selected for the routing test and should not be selected for the second-stage tests. θ is the single ability level specified for each second-stage test.

All linear programming problems discussed in this section are normally solved using the revised simplex method implemented in most computer codes. A brief description of this method and the iterative steps involved are given by Boekkooi-Timminga (1992). The computer program, OTD (Verschoor, 1991) implements the Theunissen (1985) heuristics. A prototype version of the computer program, CONTEST is currently being developed and implements the minimix and maximin models of van der Linden and Boekkooi-Timminga (1992).

CHAPTER 3

METHOD

This is a real data simulation study in which the two-stage, conventional and adaptive testing strategies are applied to item-response data obtained from the administration of a credentialling exam that had been previously administered conventionally by paper-and-pencil mode.

3.1 Data Source

Item responses to the a credentialling examination certification paper were used in this study. The exam paper consists of 250 multiple-choice items divided into 6 content subdomains in the approximate ratio of 1:2:1:1:1:1. The 3523 examinees in the response dataset were divided into 2 groups - the calibration group (1560) for the purpose of calibrating the test items and the empirical group (1934) where their actual responses were used in the simulated test administrations.

3.2 Item Pool Calibration

Item analysis was performed on the item responses based on the calibration group of 1560 examinees and 20 items with low (<0.20) or negative biserials were removed. From the 230 items in the item pool, a spaced sampling of 80 items was assessed for unidimensionality using McDonald's nonlinear factor analysis procedure enumerated in the computer program, NOHARMII (Fraser, 1989). Because of

matrix size constraints, a spaced sampling was necessary and it was assumed that the 80 items would be representative of the characteristics of the whole item pool.

The items were calibrated using the two-parameter logistic models by the computer program, BILOG 3 (Mislevy & Bock, 1989). Because of the large matrix size, it was not possible to calibrate all 230 items at one time. The response strings were divided into segments of 80, 80 and 70 items in that order, making three calibration runs in all. Item calibration took the form of a single group design in which each examinee took all 'three test forms'. No scale transformation was necessary since all examinees took the same test of 250 items at the same time.

3.3 Assessing Model-data Fit

Item response theory methodology, including its application in adaptive testing assumes unidimensionality. Dimensionality is defined as the total number of abilities required to satisfy the assumption of local independence (Lord, 1980). If a set of items is to be unidimensional, there is only one ability affecting the responses of a set of items to meet the assumption of local independence. However, in reality, several abilities unique to a few items apart from a dominant ability (ability common to all items) are possible in a set of items (Hambleton and Swaminathan, 1985; Yen, 1985). Simulation studies have shown that the dominant ability can be recovered well in the presence of minor abilities using computer programs such as LOGIST

(Reckase, 1979; Harrison, 1986). Hence, it is sufficient to show that there is one dominant ability underlying the responses to a set of items in order to apply unidimensional IRT models.

McDonald (1980,1982) developed the method of nonlinear factor analysis to account for nonlinearity of data as an improvement over linear factor analysis. This method is appropriate within the context of item response theory because the latent variable is related to performance in a nonlinear fashion. The variables in the item response model are expressed as polynomial functions of latent traits. The procedure is implemented in the computer program, NOHARMII (Fraser, 1983).

Because of matrix size constraints, a random sample of 80 items from the item pool were analyzed for dimensionality using McDonald's procedure. A response dataset is considered as essentially unidimensional if a two or more factor model do not show a significant reduction of the root mean square residuals.

Residual analysis was used to assess the fit of the 2-Parameter Model compared to the 1-P and 3-P Logistic Models.

3.4 Test Development

The responses of the empirical group of 1934 examinees were used in the simulation of test administrations based on optimal item selection. The following test designs involved the selection of test items to fit the target information

curves with test lengths and content balancing as constraints.

3.4.1 Conventional Tests

To address the first goal of the study where the improvement of test designs developed by automated techniques were compared with optimal item selection based on Birnbaum's (1968) procedure, the following conventional tests were developed:

3.4.1.1 Broad-range Conventional Tests

This test was developed to cater to a general measurement of ability over a broad ability spectrum where decision making such as grading is not important. Such tests may be used in a training program where the course instructor may need a quick assessment of the students' ability level from time to time using a short, but efficient test. The development of this test was initiated by setting a uniform target information at the ability levels: -2, 1, 0, -1, -2. The target information was set at 4.00. This target was selected based partly on the fact that the item information in the pool at the higher end of the ability continuum were rather low and a long test had to be constructed if the uniform target information was set too high. Since the abilities were transformed with a mean of 0 and a standard deviation of 1, the ability range from -2 to +2 set by the target information would have a 95% coverage of the examinees (normal distribution of the abilities were assumed) and this was considered appropriate.

The objective of the design was to create the shortest test possible that could fit into this target information. The computer program, OTD (Verschoor, 1991) was used to enumerate the design problem. The item bank file was created for input into the OTD environment. The cost function specified by the program was set to 1.00 and the content balancing option was removed. The target information of 4.00 was set from ability levels -2 to +2 in the OTD environment. Since the program made use of the normal ogive model in the item selection procedure, in order to conform to the logistic model used in this study, all a-parameters in the item bank file was multiplied by a factor of 1.7. A 486, 40 MHz computer was used for all programming work.

For the purpose of examining the efficiency of the automated procedure in optimal item selection, an optimal item selection technique, the up-and-down (UD) method (Hambleton and Swaminathan, 1985) was used. This method was chosen over the other optimal item selection methods because studies by the same researchers found that this method provided maximum information over a broader ability range (Hambleton and Swaminathan, 1985; page 252). As such, this technique would fit into the design objective where a short test was needed to cover a wide ability range. In the up-and-down (UD) procedure, an item with maximum information at ability +2 was chosen followed by items with maximum information at ability +1, 0, -1 and -2. The obtained test

information was updated each time the items were added and item selection at any particular ability level was stopped once the target information at that level was reached. The cycle was then repeated until the obtained test information had reached the target information at all specified levels. Two modifications were made to this procedure to enhance its optimal item selection. Firstly, before selecting items with maximum information at any particular ability level, a number of items with maximum information over a wide ability range was selected. This was done by computing the mean of the item information for the five ability levels and sorting the means of all 230 items in the pool from high to low. This modification was based on suggestions by Hambleton and Swaminathan (1985; page 233). Secondly, back-tracking was allowed to remove and substitute items in order to obtain the shortest test length possible and in order that the obtained test information conform as closely to the target information as much as possible. The item information matrix was computed using the software package, STATA (Computing Resource Center, 1992). All sorting and item selection were done with the aid of the software. A program was written within the STATA environment to update the test information. Since optimal item selection based on Birnbaum's procedure was done with the aid of a fast computer system, this helped speed up the item selection process.

3.4.1.2 Peaked Conventional Tests

This test was designed with the purpose of separating examinees into the pass and fail categories. The maximum amount of information was required at the region where the pass/fail decision had to be made. A criterion for passing was set at the ability level, 0.00. The test design was specified at the ability levels, -1.5, -0.5, 0.0, 0.5 and 1.5 with the target information set at 3, 10, 12, 10 and 3 respectively. The resulting test would ensure a higher precision of measurement at the region of the specified pass/fail criterion. Again, the two item selection procedures already described were used. Except for changing the shape of the target information, the OTD test design specifications were the same in the previous design for the broad-range conventional test.

3.4.1.3 Conventional Tests with Content Balancing

This conventional test was developed with content balancing as the constraint in the item selection process. The test was developed to adhere strictly to a test blueprint where a course instructor after having completed a set of instructional modules desires to have a general class assessment based on subject matter emphasis. The target information was similar to that specified in the previous conventional test. However, the test design had a fixed test length of 42 items imposed and with a content balancing constraint added. The test was developed so that the number of items selected in the six content subdomains were

6,12,6,6,6,6 in that order. This content subdomain ratio would correspond to the content specifications of the original examination paper taken by the examinees. Within the OTD environment, the content balancing constraint was fixed to take in the relative content balancing ratio. The item bank file already had the categorization of the items specified in the very beginning to indicate the stratification of the item bank.

In the use of the up-and-down (UD) method of optimal item selection, further modifications to the procedure were made. This time, instead of taking the whole item bank in the item selection process, the UD method was used for each of the six content categories and with the fixed number of items in each content subdomains in mind. The test was updated each time the first cycle of item selection was made in all six categories. Although the procedure was cumbersome, the use of the computer speeded up the process.

In addition to the above three sets of conventional tests which were designed by both the OTD and the UD procedures, an additional 42-item broad-range conventional test was designed without content balancing by OTD and was used for comparison with the 42-item content balanced conventional test to address partly the second research goal.

Four sets of comparisons were therefore made for the conventional tests:

- 1) broad-range conventional tests (OTD versus UD designs),
- 2) peaked conventional tests (OTD versus UD designs),
- 3) fixed length and content balanced conventional tests (OTD versus UD designs) and
- 4) OTD designed fixed length conventional tests (content balanced versus noncontent balanced).

3.4.2 Two-stage Tests

To address the second and third goals of the study, two forms of two-stage tests of 42-item length were developed using the automated procedure. The first form had content balancing imposed as a constraint and the second had the content balancing constraint removed in the test development process.

The target information desired for the routing tests was specified for three ability points: -1.50, 0.0 and 1.50 and the target was set at 3.00 to arrive at the optimal selection of the 14 items that will fit the target information. The optimization problem was specified in the specification file of OTD to reflect the kind of target information and the content category ratio as constraints. Once the 14 items were selected, they were removed from the item pool.

In accordance with Lord's (1980) suggestion for two-stage test development, four second stage measurement tests were developed using the automated procedure. The ability segments specified in the OTD environment for the development of these four tests were: -2.50 to -1.5, -1.5

to 0.0, 0.0 to 1.5, and 1.5 to 2.5. A uniform target was set for each ability segment and 28 items were optimally selected for each ability segment with and without content balancing.

3.4.3 Adaptive Tests

Two forms of adaptive test were developed to address further, the second and third goals of this study. A content balanced adaptive test was developed by forming item clusters or testlets from the item pool. This was done in OTD by setting target information bars of ability 1.0 in length across the ability continuum. The information bars were varied in height to adjust to the optimal selection of 7 items in a balanced content ratio and were allowed to overlap each other. Each target information bar bore the constraint of content balancing and was varied so that only 7 items were selected to form each testlet. The items were selected from the six content subdomains in the ratio 1:2:1:1:1:1. Once the items were selected, they were removed from the item pool.

The adaptive test procedure involved a search of testlets in the pool to determine which unadministered testlet had the most psychometric information at an ability level equal to a specified value. A subsample of 630 examinees from the empirical group was used for individual adaptive testing and scoring. The examinee abilities were scored by maximum likelihood procedure. The test was terminated when the variance fell below 0.10. Because of

this, the last testlet to be administered to each examinee might not be administered completely. Once the termination criterion was reached, the test was stopped, resulting in a certain number of items in the last testlet being administered instead of the complete 7 items. Content balancing was still maintained to a certain degree, although approximately.

The adaptive test procedure in which content balancing was not taken into account was the same in procedure to that described above except that instead of a search through the item pool for testlets, an item search was made without due regard to content balancing.

3.5 Scoring

The corresponding response strings of the conventional tests were created as ASCII files to serve as inputs into the MicroCAT (Assessment Systems Corporation, 1987) environment for conventional scoring. Based on the items selected for each test, the corresponding item parameters were also created as input files for the MicroCAT environment.

In the case of two-stage tests, the examinees were initially scored by the routing tests and their ability levels estimated by the maximum-likelihood procedure. Based on these initial ability estimates, the examinees were routed to their respective second stage measurement tests where their responses were scored by the same procedure.

In the scoring of the adaptive tests, an input ASCII file was created, containing the parameter values of all the items in the item bank and imported into MICROCAT (Assessment Systems Corporation, 1987). Two separate banks were set up. In the case of the content balanced adaptive test design, the items were clustered based on the testlet designs resulting from the OTD runs. In the case of the adaptive test where content balancing was not taken into account, the item pool was treated as an unstratified whole. In the administration of the testlets, the Minnesota Computerized Adaptive Testing Language (MCATL) was used to design the testing strategy which involved the following:

1. A search of the item cluster in the pool to determine which unadministered cluster had the most psychometric information at an ability level equal to a specified value.
2. The examinee abilities were scored by maximum likelihood procedure.
3. The test was terminated when the variance fell below .10.

The termination criterion corresponds to the standard error of estimate criterion of 0.3162 specified by Urry (1974) in order to achieve a fidelity coefficient exceeding .95 in simulation studies. Because of the termination criterion, the last testlet to be administered to each examinee might not be administered completely. Once the termination criterion was reached, the test was stopped,

resulting in a certain number of items in the last testlet being administered instead of the complete 7 items. Content balancing was still maintained to a certain degree, although approximately.

The test specification designed by MCATL was compiled in the MICROCAT environment. In addition to the estimated abilities, the test lengths for each examinee and the item identities were recorded. In the administration of the adaptive test without content balancing, Step 1 of the MCATL procedure was modified to an item by item search instead of searching through item clusters.

The original credentialling exam paper consisting of 250 items was taken by the examinees. After deleting 20 bad items to form the item pool, all examinees were scored on the 230 items in order to obtain ability estimates as a basis for comparison with the ability estimates obtained from the test designs. The raw scores were standardized with a mean of 0 and a standard deviation of 1.

3.6 Statistical Analysis

The item pool was assessed for unidimensionality using McDonald's procedure. The independent variables used in the study were the ability estimates from the tests developed by the various optimal item selection strategies. The dependent variable was the standardized raw scores (taken as a measure of the observed abilities) based on the examinee responses to the 230 items in the item pool.

3.6.1 Information Analysis

Data analysis began with a comparison of the target and obtained information curves. This comparison was used to address the first goal of the study. The effectiveness of an item selection procedure would be judged by how close the obtained test information was to the target information. A successfully enumerated test design problem would be shown by the obtained test information above the target at the specified ability levels with the shape of the curve as close to the target as possible. Differences in the shape of the obtained test information curves were also used to examine the effects of constraints imposed by content balancing on the test design. Computations of item and test information were done using the software package, STATA (Computing Resource Center, 1992). A computer program, INFOR was written in STATA format to compute all item and test information at various ability levels and to perform all test information plots.

The standardized raw scores of the examinees based on the 230 items in the item pool were grouped into ability groupings as follows: $-2.50 \leq \theta \leq 2.50$ at θ intervals of 0.5. If the scores cover the full range of the specified ability continuum, there will then be 10 ability groupings. Comparisons of score differences between the different test designs were based on the observed abilities.

3.6.2 Analysis of Score Differences

Two evaluative indices were used in this study. The Inaccuracy was computed as:

$$\text{INACC}(\theta) = \frac{\sum |(\hat{\theta}_i - \theta_i)|}{N} \quad (31)$$

where: N is the number of examinees in the ability grouping,

$\hat{\theta}_i$ is the estimated ability and

θ_i is the observed ability.

This index takes into account, the size of the difference between the estimated and the observed abilities. The Inaccuracies were compared between the different optimal item selection strategies.

The second index, the root mean square difference (RMSD) was computed as:

$$\text{RMSD}(\theta) = \sqrt{\frac{\sum (\hat{\theta}_i - \theta_i)^2}{n}} \quad (32)$$

This index gives more weight to larger differences between estimated and observed abilities. The computation of this statistic followed the same derivation of score differences for the RMSD. Small Inaccuracies or RMSDs will imply estimates that are closer to the observed abilities and hence, a greater level of concurrence in ability estimation for that ability grouping.

3.6.3 Correlational Analyses

Pearson product-moment correlation analyses shows the degree in which the estimated and the observed abilities go together. High correlations between scores will imply that the test strategy concerned ranks the examinees in a similar order along the ability continuum.

All computation work involving the item and test information, the RMSD, the IACC and all graphical plots were done using the software, STATA (Computing Resources Center, 1992) and the graphics and data management software, STAGE (Computing Resources Center, 1989).

CHAPTER 4

RESULTS

4.1 Unidimensionality Assessment

A stratified and spaced random sample of 80 items were drawn from the item pool to assess unidimensionality. The sampling was done to reflect the content emphasis of the six content subdomains by selecting the items in the six categories in the order: 11,25,11,11,11,11. Both linear and nonlinear factor analyses were performed on the 80 items based on 1934 examinees. A six factor solution was obtained using maximum likelihood linear factor analysis procedure implemented in the computer program, STATA (Computer Resource Center,1992). A rough approximation to unidimensionality was shown using a plot of eigenvalues of the inter-item correlation matrix. Figure 1 shows the dominant first factor and a high ratio of the first to second factor eigenvalues, which is a rough indication of unidimensionality (Reckase, 1979). Table 1 shows that the percentage variance accounted for by the first factor was very high in comparison with the other factors.

One to six factor models were specified in McDonald's nonlinear factor analysis procedure and enumerated in NOHARMII (Fraser, 1989). Results of the analysis showed that for two or higher factor models, the mean square residuals did not improve very much over that of the

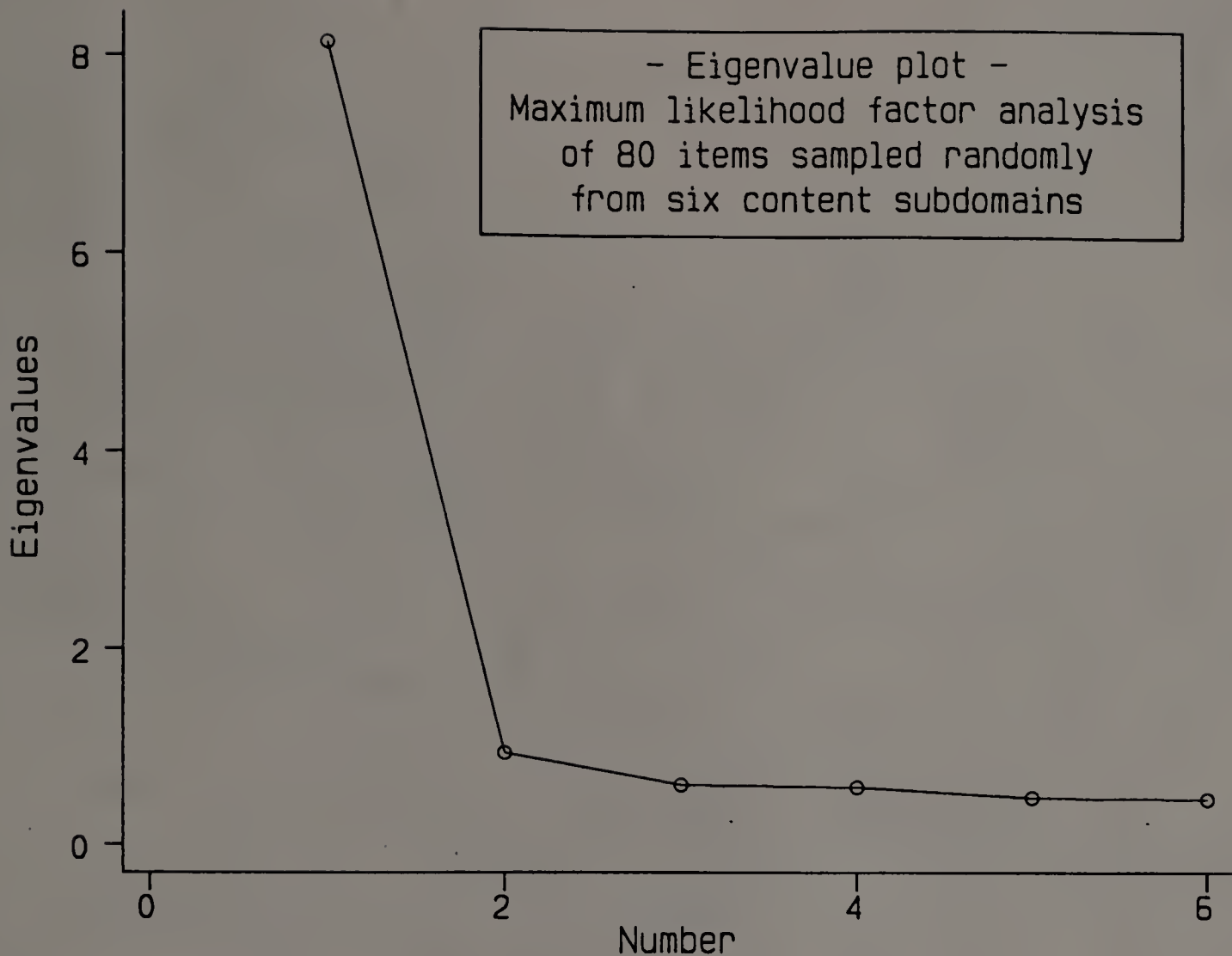


Figure 1. Plot of Eigenvalues of Inter-Item Correlation Matrix

the one-factor model. The degree of improvement was only about 2.0% for the six-factor model (see Table 1). The item pool was deemed to have essential unidimensionality, a condition fulfilled for application of IRT in testing.

Model-data fit was assessed using residual analysis in a previous study (Hambleton, Dirir & Lam, 1992). Table 2 shows that 11.9% of the absolute valued standardized residuals exceeded a value of 3 when the 1-p logistic model was fitted to the data. The residuals between 2 and 3 for the same model exceeded that of the normal distribution by a factor of 3 indicating that the 1-p logistic model showed

Table 1

Fit statistics for linear and nonlinear factor models

Model	Fit indices			
	λ_1	% Var	SS _{res}	MS _{res}
Linear factor analysis				
Factor				
1	8.16	10.2		
2	0.94	1.2		
3	0.61	0.8		
4	0.59	0.7		
5	0.48	0.6		
6	0.47	0.6		
Nonlinear factor analysis				
1-factor			0.0713	0.0475
2-factor			0.0702	0.0472
3-factor			0.0702	0.0471
4-factor			0.0683	0.0466
5-factor			0.0676	0.0462
6-factor			0.0685	0.0466

the poorest fit. The distributions of the residuals for the 2- and 3-p logistic models were quite close to each other while the residual distribution of the 3-p model approximated closest to the normal distribution. Although the 3-p logistic model fitted the test data best, the 2-p model was used in order to accommodate the version of OTD software which did not cater for the 3-p logistic model.

4.2 Descriptive Statistics

Descriptive statistics of the item pool showed that the items were generally easy and differed in discrimination in the content subdomains (see Table 3). Items in Subdomain F

Table 2

Analysis of Standardized Residuals for the 1-, 2- and 3-Parameter Logistic Models

Logistic Model	0 to 1	1 to 2	2 to 3	> 3
1	44.1	30.5	13.5	11.9
2	61.6	30.1	6.1	2.3
3	66.5	26.5	5.7	1.3
Normal Dist.	68.2	27.2	4.2	0.4

generally have higher discriminations and items in Subdomain A are very easy compared to the rest. The differing characteristics of the item parameters may have a bearing in the optimal item selection process as can be seen later.

4.3 Comparison of OTD and UD Designed Broad-range Tests

Figure 2 shows the obtained information functions of both conventional tests developed by the binary programming (OTD) procedure and the modified up-and-down (UD) optimal item selection procedure. Successful enumeration of the optimization problem with uniform target set at 4 from ability -2 to +2 resulted in the selection of a minimum of 30 items. The UD method on the other hand, resulted in a selection of 35 items. The obtained information functions were very close at the higher end of the ability continuum but differed greatly at the middle portion of the ability

Table 3

Distribution and Descriptive Statistics of Test Items
by Content in Item Pool

Content	Number of items	Parm.	Descriptive statistics			
			Mean	S.D.	Min	Max
A	27	a	0.49	0.16	0.23	0.91
		b	-0.58	1.04	-3.20	1.54
B	78	a	0.55	0.17	0.22	1.00
		b	-0.24	0.89	-2.34	2.35
C	31	a	0.46	0.13	0.24	0.73
		b	-0.21	1.09	-2.82	2.69
D	30	a	0.48	0.14	0.26	0.70
		b	-0.28	0.95	-2.06	1.89
E	27	a	0.58	0.15	0.38	0.88
		b	-0.17	0.88	-1.96	1.80
F	37	a	0.64	0.21	0.28	1.05
		b	-0.22	0.51	-1.19	0.74
Item bank	230	a	0.54	0.17	0.22	1.05
		b	0.27	0.90	-3.20	2.69

continuum. Item selection by the automated procedure appeared to have the advantage of improving on the obtained information function compared to the manual procedure by approximating closer to the target information. The automated test procedure also resulted in the development of a shorter test compared to the manual procedure.

4.4 Comparison of OTD and UD Designed Peaked Tests

Figure 3 shows the obtained test information curves of the peaked tests developed by the OTD and the UD procedure. This time, the two curves were very close, indicating that, with a peaked target, the automated procedure appeared to have a smaller improvement over the manual procedure. Both

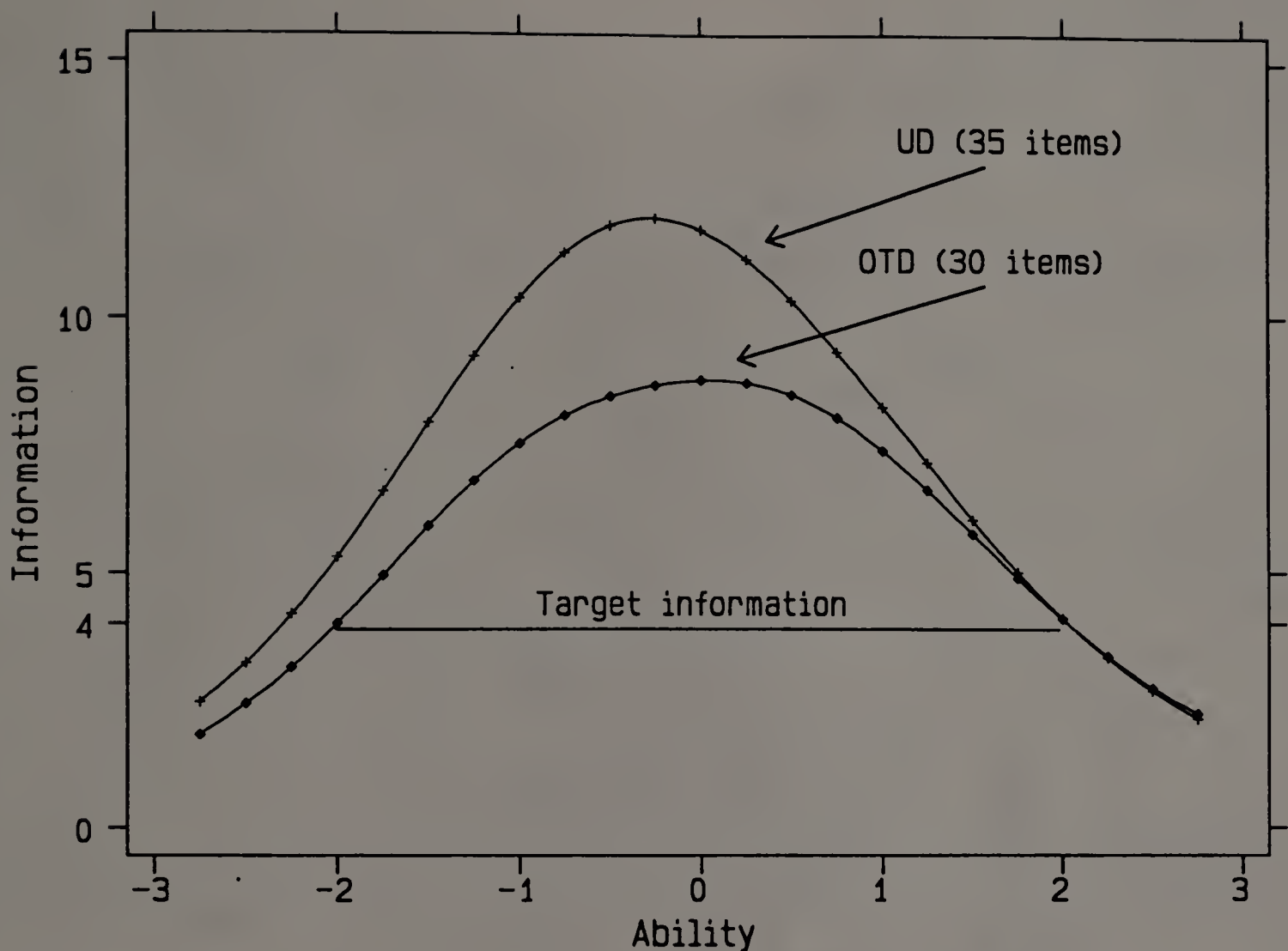


Figure 2. Obtained Test Information Functions of UD and OTD Designed Broad-range Conventional Tests

curves were shifted to the left of the target information because of the greater distribution of items with larger information at the lower ability levels. However, the obtained information curve arising from the manual procedure was shifted further to the left indicating a lesser approximation to the target information at the lower end of the ability continuum. Test lengths from both item selection procedures were almost the same.

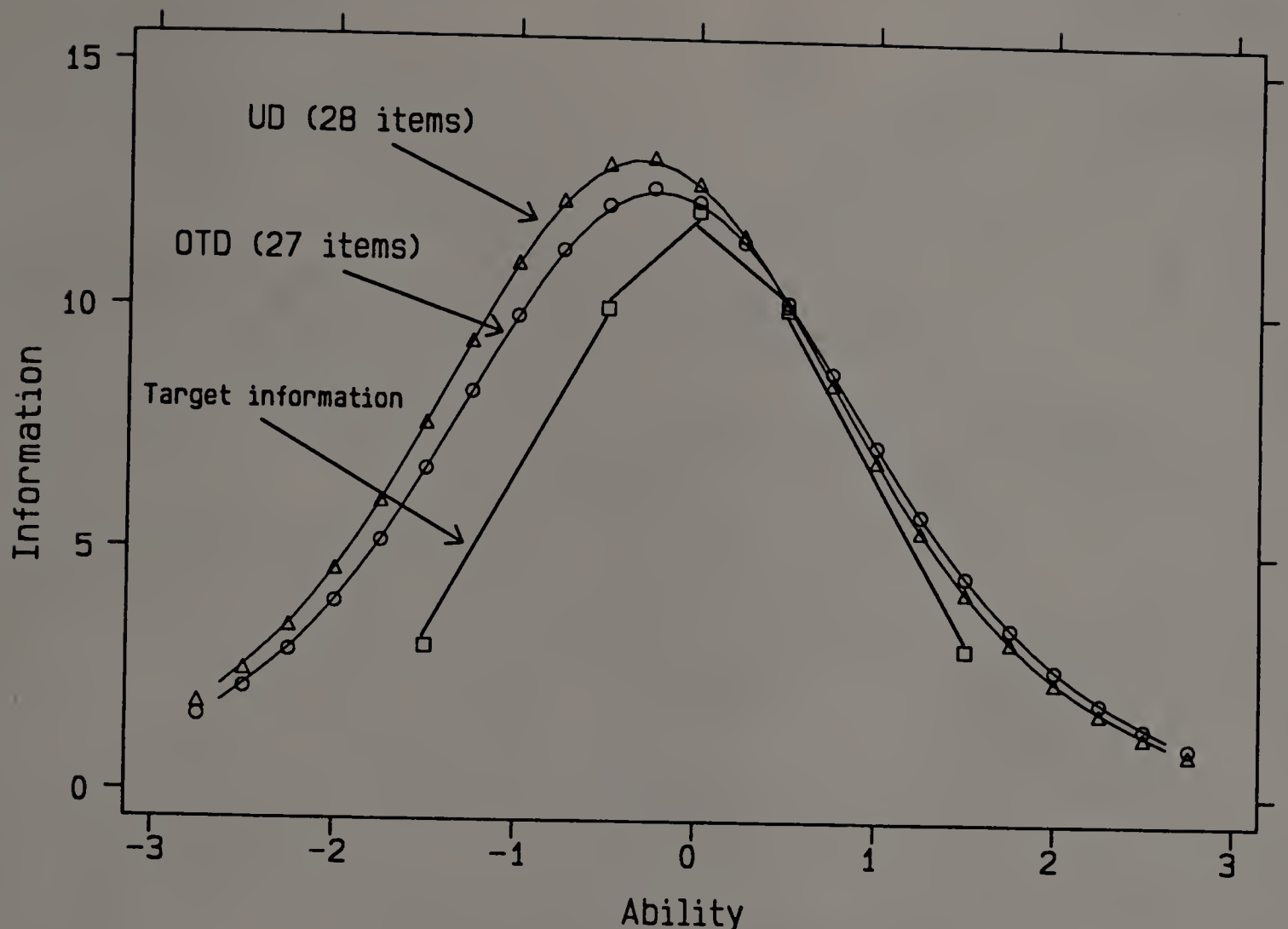


Figure 3. Obtained Test Information Functions of UD and OTD Designed Peaked Conventional Tests

4.5 Comparison of Content Balanced Conventional Tests

Figure 4 shows the obtained test information of two conventional tests with the constraint of content balancing and a fixed test length of 42 items imposed. The design which used the UD procedure was not quite successfully enumerated at the ability level of +2. The obtained test information was slightly below the target information at this ability level (see Table 4). Again, at the lower end of the ability continuum, the manual procedure of item selection showed a lower approximation to the target

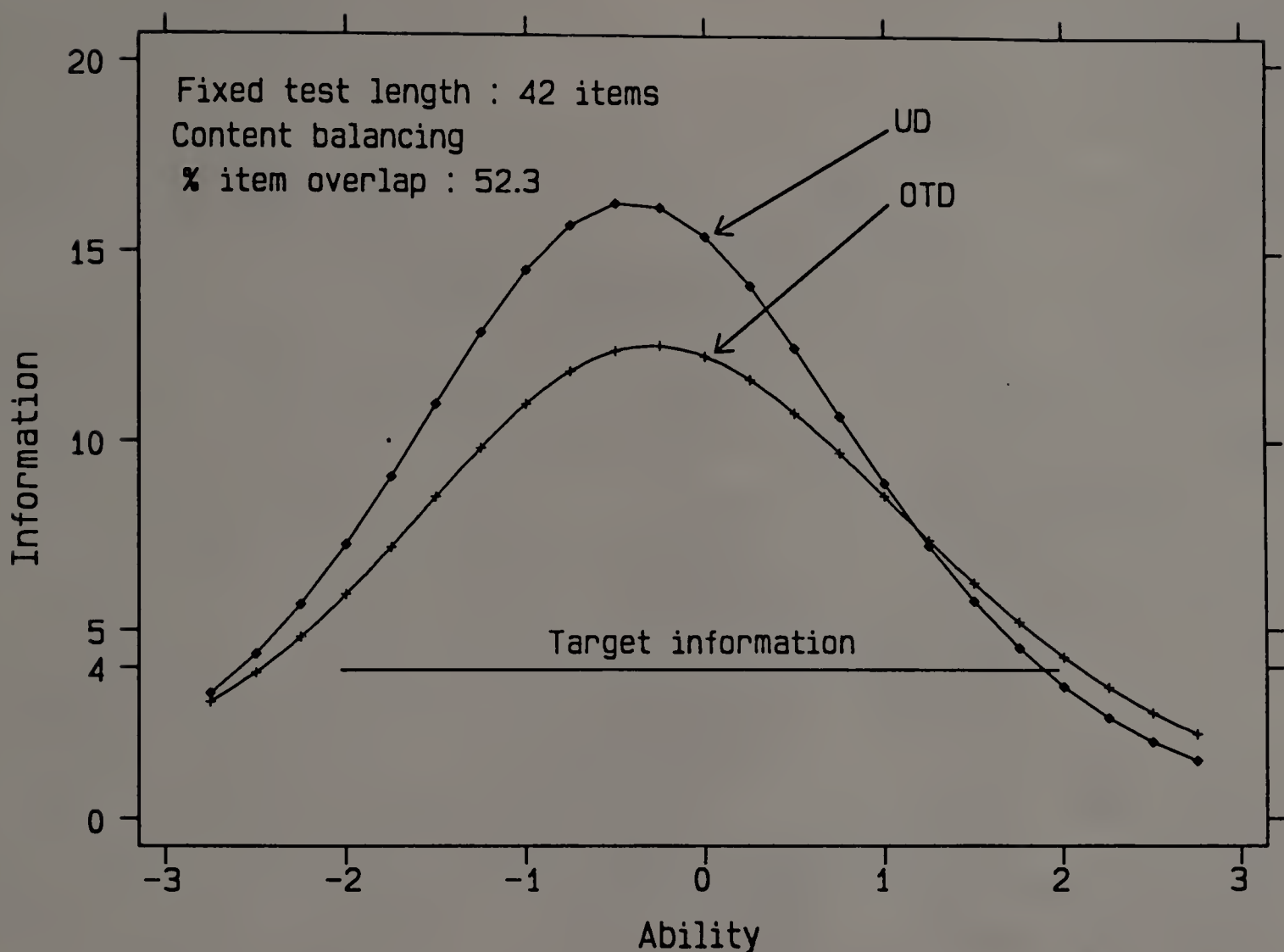


Figure 4. Obtained Test Information Functions of UD and OTD Designed Conventional Tests with Content Balancing

information. The percentage of item overlap for the two item construction procedures was 52.3.

4.6 Comparison of Tests with Content Balancing Constraint

Figure 5 shows the obtained test information curves for two fixed length conventional tests of 42 items developed by OTD. One test had the content balancing constraint imposed and the other had the constraint removed. With test length and target information held constant for both designs, the measurement precision of both tests can be examined by comparing both obtained test information curves. It can be

Table 4

Obtained and Target Information Functions of Specified Ability Levels for Conventional Test Designs

Test design	Ability Level					Test Length
	-2.00	-1.00	0.00	1.00	2.00	
Broad-range test						
Target	4.00	4.00	4.00	4.00	4.00	
OTD	4.02	7.53	8.76	7.39	4.10	30
UD	5.32	10.39	11.71	8.24	4.12	35
Peaked test						
Target	2.00	10.00	12.00	10.00	2.00	
OTD	3.92	12.11	12.20	10.15	2.60	27
UD	4.58	12.96	12.58	10.10	2.35	28
Broad-range test with content balancing						
Target	4.00	4.00	4.00	4.00	4.00	
OTD	5.94	10.97	12.22	8.53	4.28	42
UD	7.26	14.49	15.37	8.88	3.50	42

seen that the test with the content balancing constraint removed has a higher information in the middle range of the ability continuum. A possible explanation of the differences in the test information is that the imposition of a content balancing constraint in the test design resulted in the forced selection of items of lesser information across content subdomains in order to fulfil the content ratio specification of the six content subdomains. The uneven distribution of the items in the six content subdomains could be seen when content balancing was lifted

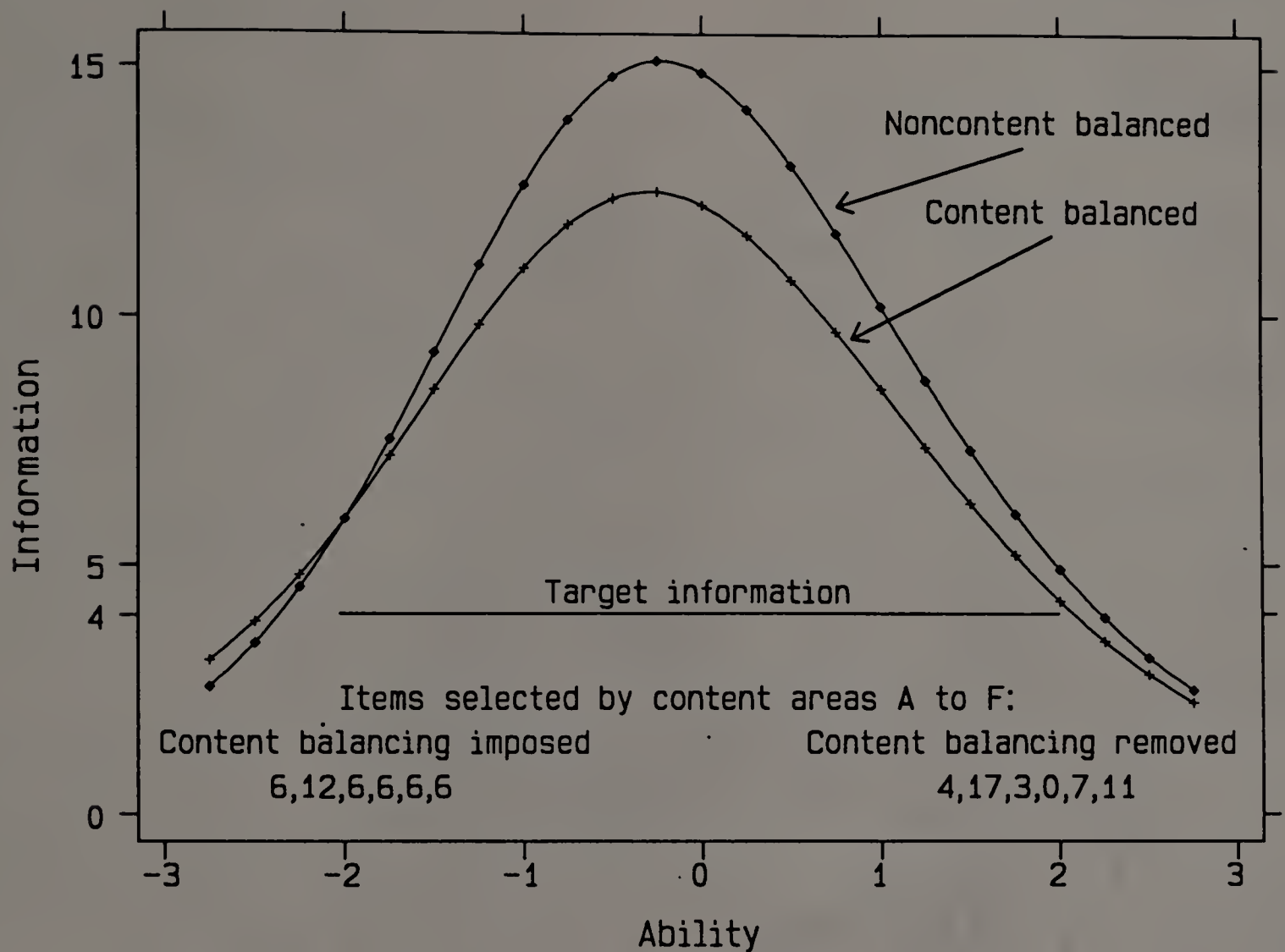


Figure 5. Obtained Test Information Functions of OTD Designed Conventional Tests with and without Content Balancing

(see Figure 5). More items from Subdomains B and F were selected at the expense of the other content areas. No items were selected from D.

Table 5 shows the correlation of the ability scores of the eight conventional tests with the observed abilities based on the 230-item bank.

The examinees were scored on the two conventional tests and the INACCs and RMSDs (both previously defined) were computed for 9 ability intervals (no examinees were found in

Table 5

Correlation of Conventional Test Scores with Standardized
Raw Scores

Broad-range tests	Standardized Raw Scores
1) OTD designed (30 items)	0.92
2) UD designed (35 items)	0.92
Peaked tests	
3) OTD designed (27 items)	0.90
4) UD designed (26 items)	0.89
Fixed length content balanced (42 items)	
5) OTD designed	0.93
6) UD designed	0.93
Fixed length OTD designed (42 items)	
7) Content balanced	0.93
8) Noncontent balanced	0.94

N = 1934

the ability interval from 2.00 to 2.50) (see Figures 6 and 7). These score differences were based on the criterion abilities estimated from the 230 items of the item pool. Although the INACCs and RMSDs of the noncontent balanced test were slightly lower than those of the content balanced counterpart, their differences were not so significant. Both MAD and RMSD were seen to increase towards the higher end of the ability continuum.

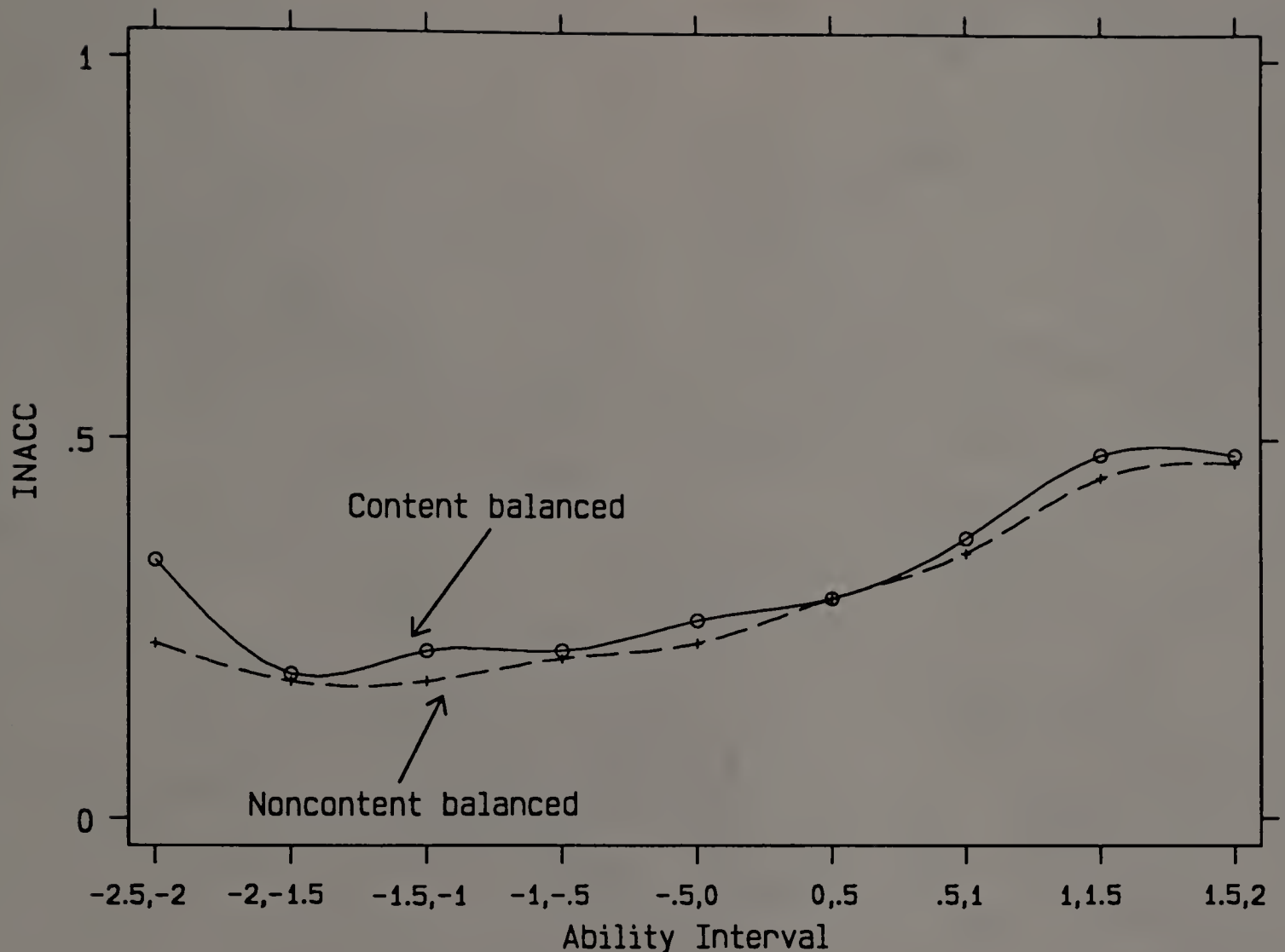


Figure 6. INACC Plots for Conventional Tests with and without Content Balancing

4.7 Comparison of Two-stage Test Designs

In the two-stage test designs in which the examinees were routed to the respective second stage measurement tests by the routing test, the INACC and RMSD are almost identical in the middle section of the ability continuum (see Figures 8 and 9), indicating the efficiency of the routing test in correctly channelling the examinees to the respective second stage test. At the higher and lower ability levels, the INACC and RMSD differences between the two designs differed, indicating a greater loss of measurement precision for the

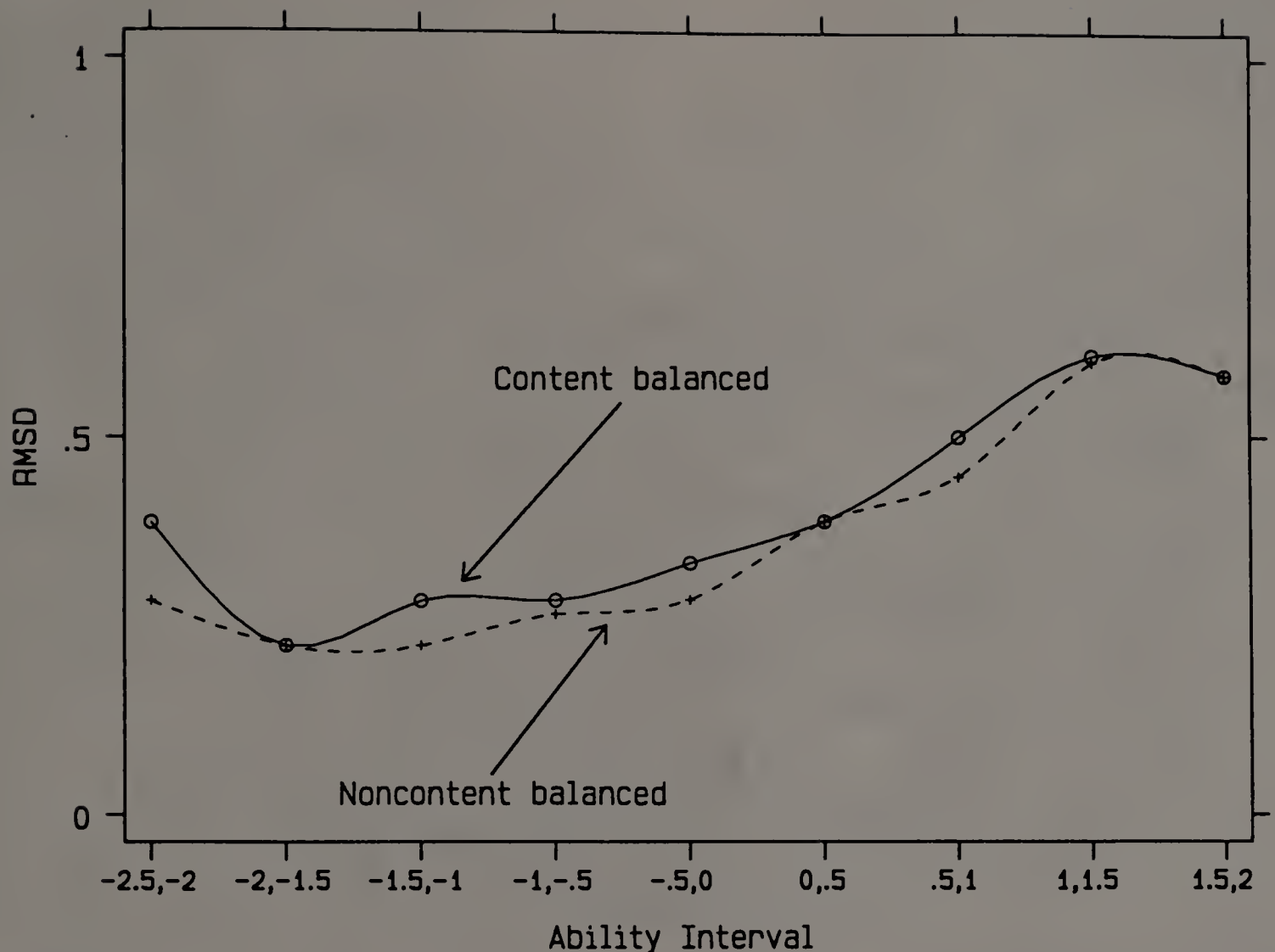


Figure 7. RMSD Plots for Conventional Tests with and without Content Balancing

content balanced conventional test. The dip in the INACC and RMSD was seen at the high extreme end of the ability continuum indicating that the two-stage test was doing its job of measuring more precisely at the extreme ends of the ability continuum. Hence, the two-stage test showed an improvement in measurement precision over that of the conventional tests in this regard.

The correlation of the test scores from the content balanced and from the noncontent balanced two-stage tests with the observed abilities were 0.89 and 0.91 respectively,

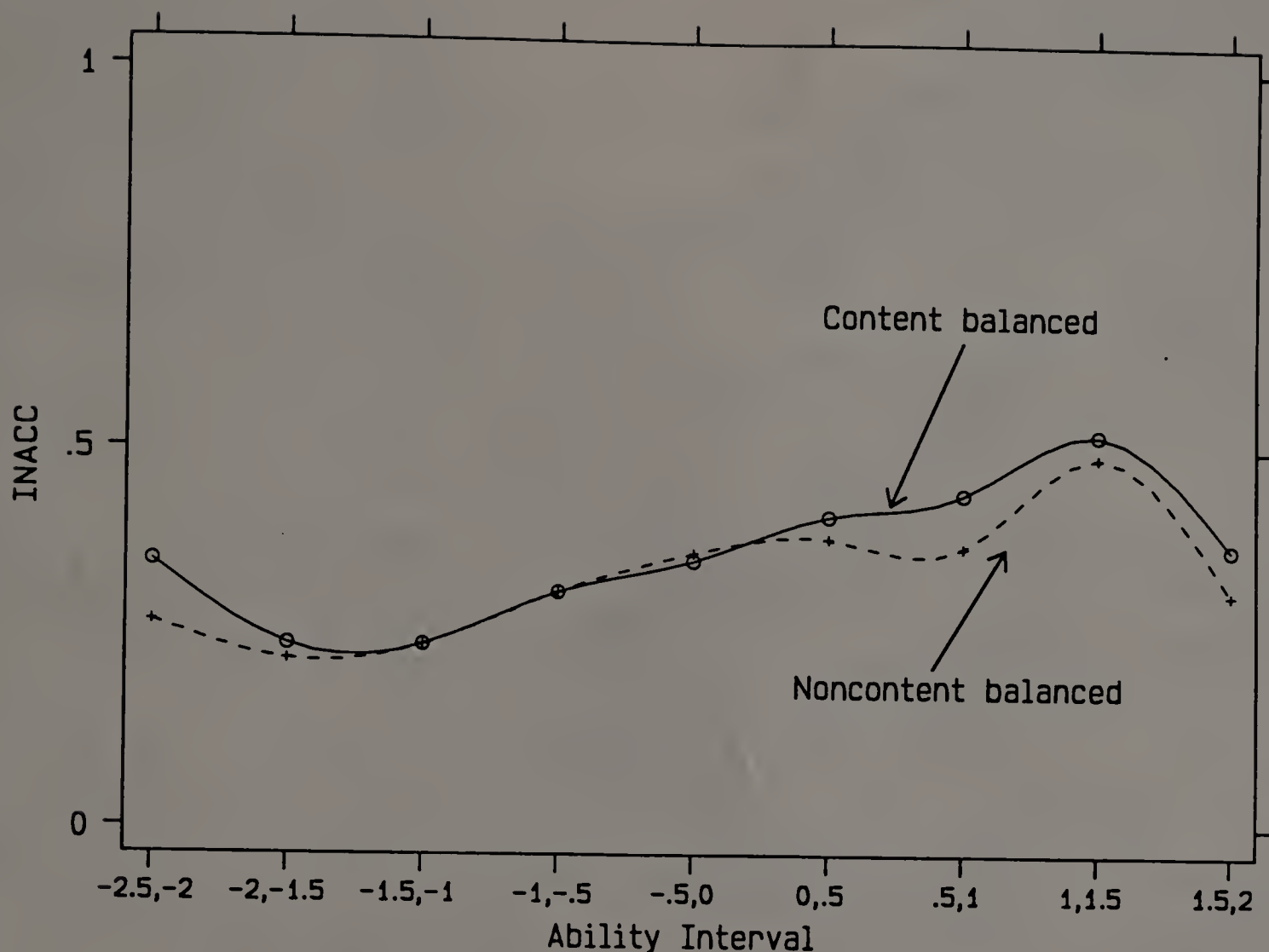


Figure 8. INACC Plots for Two-stage Tests with and without Content Balancing

indicating that both test designs did not differ very much in recovering the criterion abilities. Both the conventional tests and the two-stage tests showed limitations in that the INACCs and the RMSDs were relatively high especially towards the higher ability levels.

4.8 Comparison of Adaptive Test Designs

Figure 10 shows the result of using two target information bars to optimally select testlets of 7 items each to form the content balanced adaptive test. A total of 24 testlets were formed. The remaining items could not be

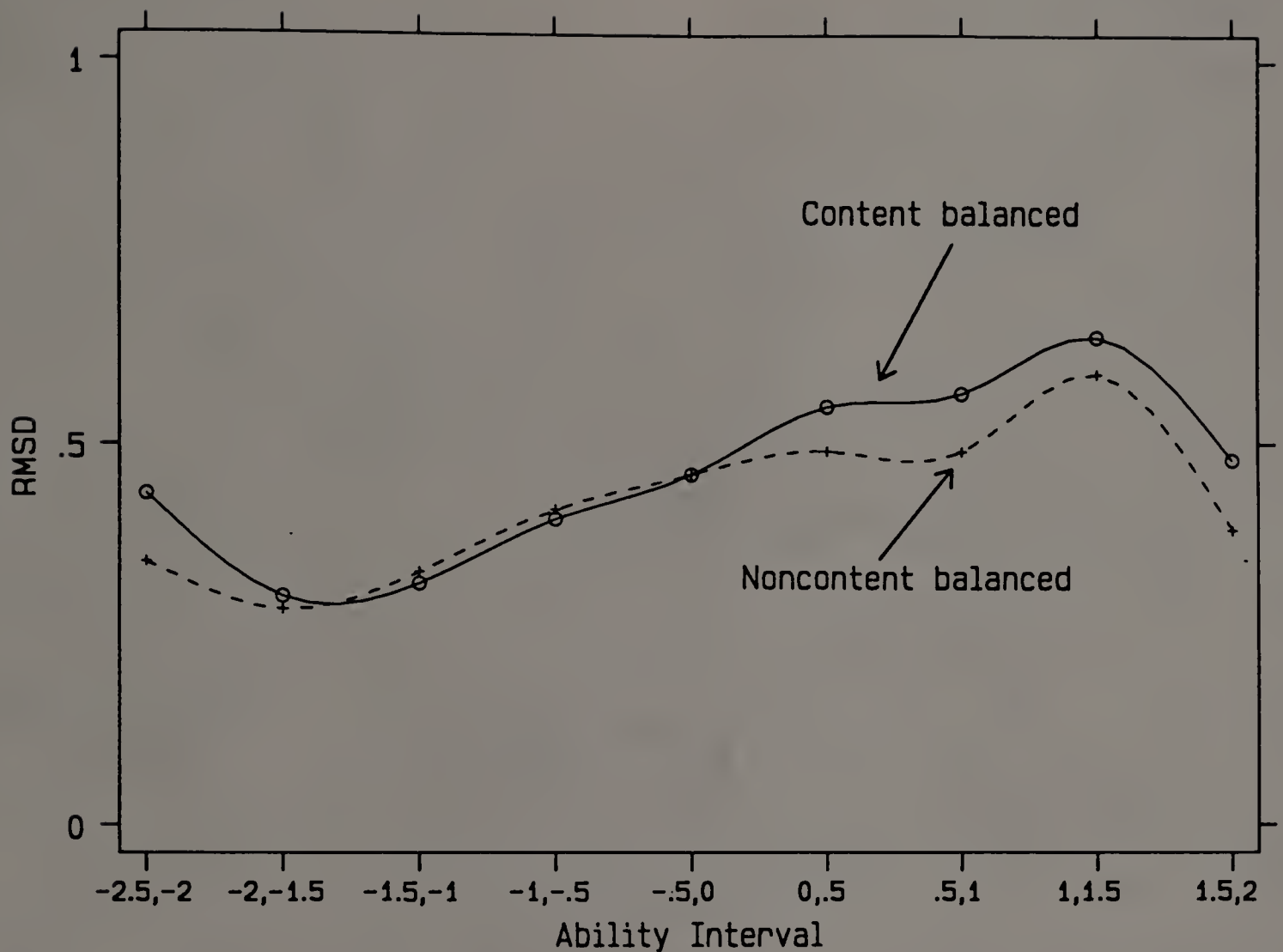


Figure 9. RMSD Plots for Two-stage tests with and without Content Balancing

successfully selected to fit the information, even though lowered to the minimum and OTD failed to enumerate the problem each time. As such the best 148 items were clustered to form the content balanced adaptive test item pool. The testlets comprised items bearing the same specified subdomain ratio that is 1:2:1:1:1:1 in the six subdomains in that order. In each testlet, the items were arranged in the order: A,C,D,E,F,B to maintain consistency throughout the test administration process. As in the case

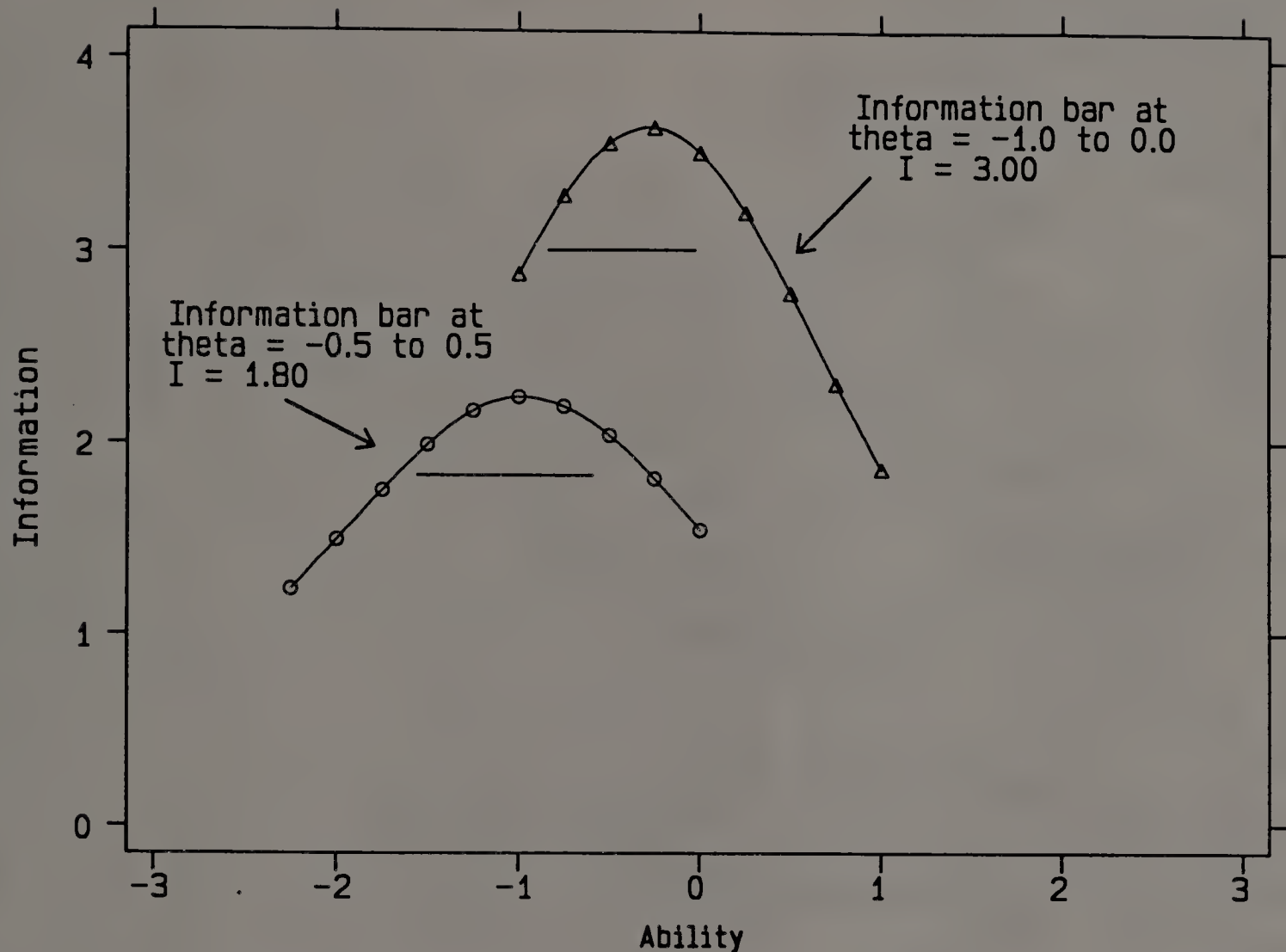


Figure 10. Testlet Target Information Bars

of the routing test, the testlets represented little peaked tests gleaned from the whole item pool.

The mean item length, minimum and maximum number of items administered for the content balanced adaptive test were 35, 24 and 90 in that order. For the adaptive test in which content balancing was not considered, the values were 30, 21 and 60 in that order. The longer test administration for some examinees was an indication of convergence difficulties probably due to some aberrant responses since a real dataset was used.

Figures 11 and 12 show plots of the INACCs and RMSDs of the two adaptive test designs. Because of the nature of the sampling, no examinees were found with observed abilities lower than -1.5 and more than 2.0. The plots were observed to be consistent throughout the whole ability levels, especially at the extreme ends, indicating almost similar measurement precision across abilities which is a feature of adaptive testing. What was most significant was that both INACCs and RMSDs were lower than those of the conventional and two-stage tests which indicate a further improvement in measurement precision especially at the extreme ends of the ability continuum. However, the score differences were higher in the content balanced adaptive test in some sections of the ability continuum and lower in the other sections of the ability continuum. The correlation with the observed abilities for the content balanced and the noncontent-balanced adaptive tests were 0.90 and 0.91 respectively.

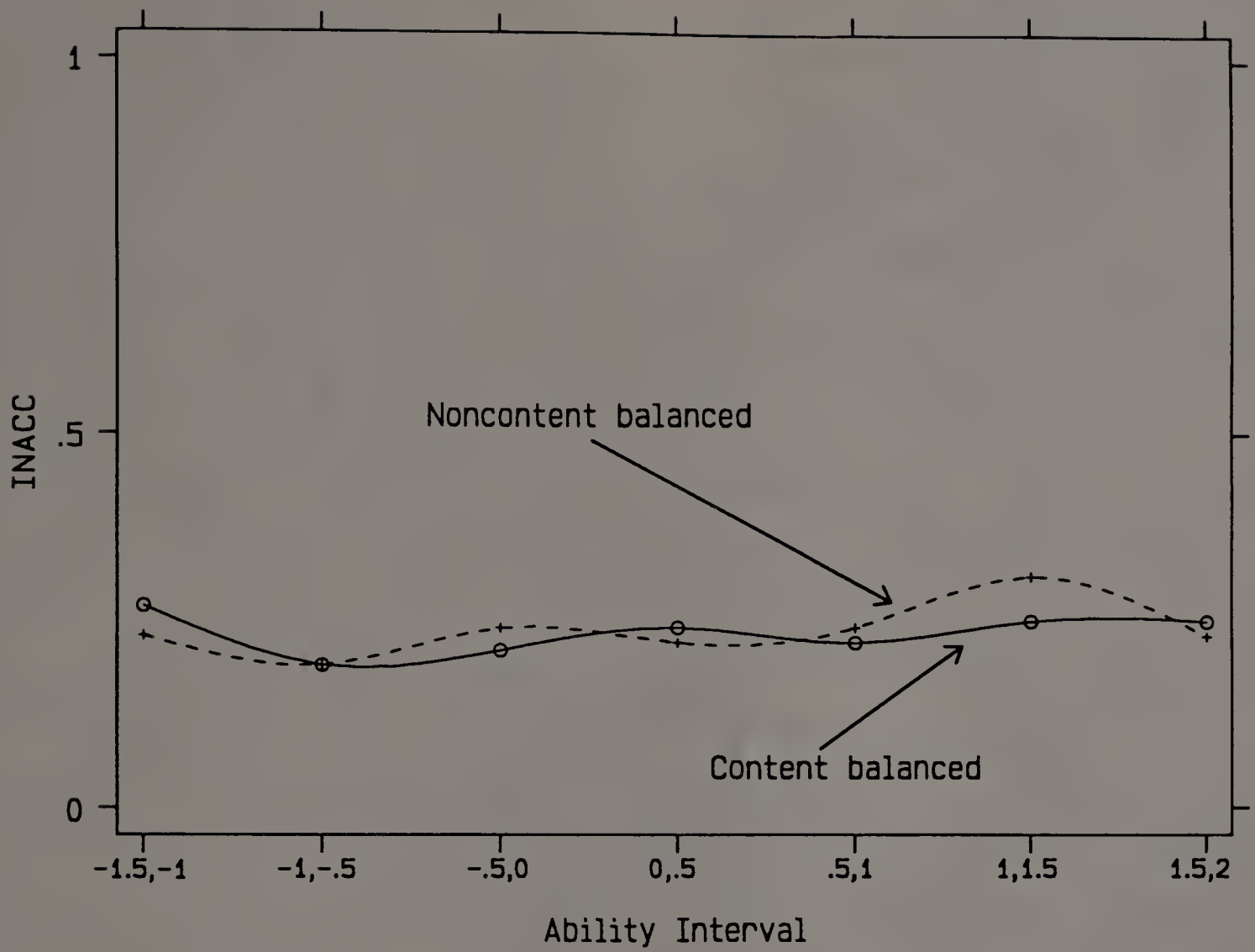


Figure 11. INACC plots for Adaptive Tests with and without Content Balancing

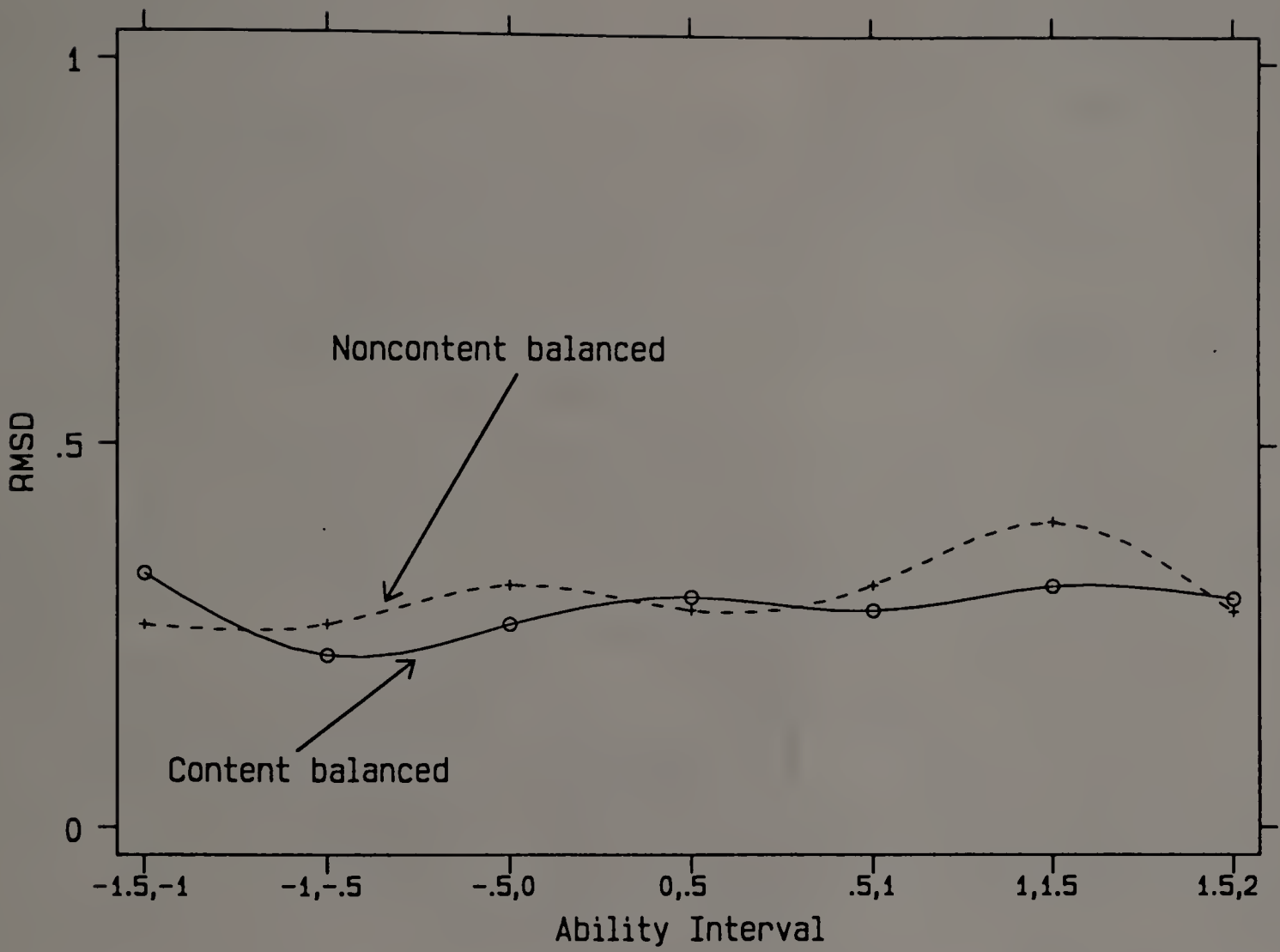


Figure 12. RMSD Plots for adaptive Tests with and without Content Balancing

CHAPTER 5

DISCUSSION AND CONCLUSION

Though the concept of optimal item selection and target information curve fitting is nothing new and had been introduced as far back as 1965 by Lord, the method involved in the past appeared to be somewhat rigorous and was based on a manual item by item selection procedure with the construction of the test information curve each time to examine its fit with the target information, as outlined by Lord (1980) and implemented in a set of heuristics by Hambleton and Swaminathan (1985). The test constructor is guided by the kinds of item difficulties and discriminations needed to fill the gaps that need to be filled in order to fit the target information curve. This manual procedure though somewhat rigorous, no doubt gives the test constructor a hands-on experience of seeing the change in test information as items are added or deleted. As such, the test constructor is fully in control of the test development process. The heuristics used by Hambleton and Swaminathan (1985) and modified in this study gave a systematic way of optimal item selection. With the help of the computer in performing all the computation and plotting work, the heuristics could be implemented fairly easily.

5.1 Conventional Test Designs

For all the conventional test designs, the manual up-and-down (UD) procedure took about 30 minutes to enumerate

each design problem while the automated procedure took only a few seconds when a 486 machine was used. The results showed in general, that the automated procedure based on the binary programming approach showed a closer approximation to the target information compared to the manual procedure of optimal item selection when a uniform target was used. Where a fixed test length was not imposed, the automated procedure produced a shorter test. With a peaked target, the improvement made by the automated test design over the manual UD procedure was not so apparent. Both methods also revealed difficulties in fitting the target information function towards the lower end of the ability continuum because of the higher distribution of easier items compared to more difficult items. The study showed that the manual procedure of optimal item selection yielded results that were almost as good as the tests developed by OTD. This could be seen by the closeness of the obtained test information curves and the high correlations between the estimated abilities with the criterion abilities (see Table 5, Chapter 4).

The results also showed the efficiency of binary programming which attempts to select the best and optimal items despite the content-balancing constraint. On the other hand, the manual UD procedure could also approximate the results fairly closely despite the complexity of cycling the procedure across content subdomains. Problems highlighted by Boekkooi-Timminga (1992) concerning the

difficulties of arriving at optimal solutions especially when additional content-balancing constraint was added could be minimized since back-tracking work can be quickened by the computer. Although the percentage overlap of items between the two conventional tests with fixed test lengths and content-balancing, was only 52.3, the correlation of scores between the two tests was 0.93. This is an indication of the property of the item bank where the IRT assumptions are met in which the items are fungible (interchangeable) units. At the item selection level, the study illustrated the sensitivity of content-balancing. The distribution of items differed significantly between the content-balanced test and a test without content-balancing.

5.2 Two-stage Test Designs

In two-stage tests, the difficulty in optimal item selection could be seen at the extreme ends; in particular, the higher end of the ability continuum because of the lower distribution of more difficult and discriminating items. Where the distribution of good items is high as in the middle region of the ability band, almost equiprecise measurement were found between the two test designs. Because of the relatively greater number of good items across content subdomains around the middle region of the ability continuum, test information between the content-balanced and noncontent-balanced designs did not differ very much even though item selection combination differs. In other words, imposing a content-balancing constraint would

not affect measurement precision significantly provided that there is a good distribution of items across the content subdomains. This is also true of conventional tests.

5.3 Adaptive Test Designs

The effect of adaptive tests in lowering the score differences and ensuring equiprecise measurement have been shown in both content balanced and noncontent balanced designs. The use of OTD designed testlets for incorporation into a testlet based adaptive test system is an improvement over Kingsbury and Zara's (1992) testlet adaptive test model based on item selection by clustering of item difficulty levels. While the Kingsbury and Zara's model showed significant differences in measurement precision between the content balanced and the noncontent balanced adaptive tests, the model used in this study resulted in narrowing the gap between the two designs even though the item pool is less than ideal when compared to a simulated item pool used by the researchers. In the model used in this study, the small INACCs and RMSDs between the two adaptive test designs is an indication of the efficiency of automated item selection in selecting optimally, the items across content subdomains. An interesting part of the results was that in many sections of the ability continuum, the score differences of the content balanced adaptive test were actually lower than the noncontent balanced counterpart. This could partly be explained by the efficiency of both OTD and the adaptive algorithm in selecting the best items within the constraints

of content balancing to the extent that it even improved upon the adaptive test without the content balancing constraint. The results of the OTD designed testlet form of adaptive testing are very encouraging and indicates the viability of developing an efficient content balanced adaptive test.

Finally, it must be noted that although the correlations between the estimated abilities from the test designs and the criterion abilities are high, the INACCs and RMSDs were different across test designs. This is because the correlation coefficient is a measure of how two sets of scores go together but the INACCs and the RMSDs are measures of how close the test scores are with the standardized raw scores. The high correlations between the test designs with different item combinations and the criterion scores are also a good indication that the assumption of IRT concerning item fungibility are met.

5.4 Possible Application of Automated Test Designs in the Schools

In the Singapore situation where every school is fully equipped with the necessary computer hardware and with sufficient government funding for the purchasing of software, the use of automated test development is a viable option. This is because of the availability of a core of teachers trained in basic test theory and the availability of items banks which are centrally linked to the Ministry of Education.

Because of the policy of continuous assessment in schools for diagnostic testing, streaming and for promotion purposes, the use of automated test designs will speed up the process of test development. The current practice among test practitioners is to select the items from the Ministry of Education central item banks based on the test specification tables. The items are selected based on the classical criteria of p-values between 0.4 and 0.6 although items are already calibrated using the Rasch model and in many cases, using the 2- and 3- parameter models. This apparent discrepancy in such item selection procedures stemmed from the difficulties and the time involved in applying Birnbaum's IRT-based methods of test construction. It is possible in the near future for teachers to improve the test development process by making use of automated methods of test designs using OTD or the yet to be released, CONTEST.

The use adaptive testing as a form of continuous and diagnostic assessment together with the aid of OTD for content balancing will assure the school administrators that test specifications will be adhered to and give better credibility to the use of adaptive procedures. The tradeoff of course, will be a longer adaptive test in order to adhere to the constraints of content balancing.

5.5 Conclusion

Content specification is one of the important procedures to be followed in many school-based assessment

programs. This is also true in licensure and certification programs. While adaptive testing has been in use for a good number of years, one of the many concerns prior to its acceptance is the need for content balancing. Apart from content balancing, the test blueprint may also require the balancing of item format as well as balancing the skill levels tested by the items. These added variables will impose a heavy load on the manual procedure especially if the stratified item pool is large in order to accommodate the different item categories.

The study shows that content balancing in test designs using binary programming procedures in OTD was done without the significant expense of measurement precision. Automated test designs used in an adaptive testing environment in a modified testlet based model reduces any possible loss of measurement precision even though the distribution of item parameters across content subdomains is uneven. It could also be seen that in a real item pool where the item discriminations are generally smaller when compared to those generated by the computer in simulation studies, the item information curves would be generally flatter and the use of OTD in this connection, would be an advantage in terms of efficiency and time.

The application of linear programming in test designs as implemented in the computer program, OTD is a viable option and have been shown to improve the results of a test designs. This method involved setting a target information

and assured the test developer that the test would conform to a certain level of precision. This is of particular advantage especially when dealing with a less than ideal item pool where the distribution of item difficulties and item discriminations differ across content subdomains. However, the use of a real item pool is more realistic and reflects the problems associated with test designs in the real world. However, one must bear in mind the limitation of using OTD. Because of the binary programming algorithm used by Theunissen (1985) and implemented in OTD, the resulting test information always has a characteristic hump even though the target information is uniform. Nevertheless, the use of OTD as against the manual UD procedure is still an advantage as the obtained test information curves using OTD were significantly lower for uniform targets.

Finally, although automated test designs offer the ease and efficiency in which a test is built by the computer, the test developer is still in control. Current software technology does not account for cross-item clue elimination. Hence the need for the test developer to ensure that this procedure is enforced especially when dealing with a large item pool. The test developer will also need to ensure the correct sequencing and layout of the test items forming the test.

5.6 Suggestions for Further Study

It is envisaged that discrepancies in test information when content balancing constraint is imposed can be reduced if distribution of item characteristics across content subdomains is homogeneous. This can be done by increasing the number of items in each content subdomain to reflect a homogeneous measurement precision across content subdomains. Alternatively, the content subdomains can be collapsed to a smaller number. Further study in the application of optimal test design procedures needs to be looked into when these are taken into consideration. The practical implication at this point is that there is a need for any good item bank to constantly upgrade its pool especially when stratification is involved.

As already pointed out by van der Linden (1987), the binary programming model used by Theunissen (1985) and implemented in OTD resulted in a characteristic hump in the obtained test information function even though the target was set to be uniform across ability levels. This is because of the way in which the algorithm will select more items located in the middle of the interval specified by the target, resulting in a high test information in this region. As such, it is near impossible to develop a rectangular test with equiprecise measurement across ability levels using OTD. However, the study indicated that with the use of a peaked target, OTD appeared to handle the optimal solution very well. The minimax and maximin models developed by van

der Linden (1987) specify the minimization of the largest deviation between the test information and the target information and result in a closer approximation to the target. The prototype software, CONTEST (van der Linden, 1992) was recently developed to handle this model. An area for further study will be a comparison of the efficiencies of van der Linden's (1987) minimax/maximin models and Theunissen's (1985) model in optimally selecting items given a uniform, peaked and bimodal target and the implications of these models in test construction.

Finally, although the study indicated the success for the use of OTD designed testlet form of adaptive testing, no comparison was made with other forms of content balancing methods in adaptive testing. One possibility for future research could be a comparative assessment of different forms of content balancing in adaptive testing that includes the Kingsbury and Zara's (1992) constraint and testlet forms of adaptive testing with the OTD-testlet procedure.

APPENDIX

ITEM BANK PARAMETERS

ITEM BANK PARAMETERS

Item	a	b	Content
1	0.54	-1.41	11
2	0.54	-0.74	12
3	0.58	-0.36	12
4	0.91	-0.82	11
5	0.55	-0.17	14
6	0.61	-0.87	14
7	0.41	-2.30	11
8	0.67	-1.01	12
9	0.38	-2.34	12
10	0.52	0.45	11
11	0.45	0.72	15
12	0.38	-0.00	16
13	0.68	1.32	12
14	0.41	-0.90	16
15	0.37	0.11	16
16	0.32	0.39	16
18	0.34	0.09	12
19	0.38	0.41	12
20	0.40	0.72	12
21	0.64	0.01	12
22	0.50	0.35	12
23	0.52	0.14	12
24	1.01	-0.32	12
25	0.47	0.77	12
26	0.69	0.57	15
27	0.77	-0.19	15
28	0.88	-0.92	15
29	0.63	0.69	15
30	0.52	-0.69	11
31	0.42	0.14	12
32	0.41	1.23	15
34	0.36	-2.61	13
35	0.57	0.48	13
36	0.51	0.34	13
37	0.84	-0.22	12
38	0.68	0.04	14
39	0.70	0.81	12
40	0.30	0.92	12
41	0.59	0.39	12
43	0.33	-0.24	11
44	0.66	-0.25	15
45	0.63	0.45	15
46	0.53	-0.72	11
47	0.50	-1.11	11
48	0.33	-0.99	12
49	0.30	-1.33	12
50	0.70	-0.46	14
52	0.60	0.46	11
53	0.34	0.61	11
54	0.47	-0.69	12

55	0.52	-1.19	11
56	0.53	0.28	15
57	0.43	1.36	15
58	0.67	-0.60	12
60	0.42	-1.03	14
62	0.42	0.47	14
63	0.65	0.52	12
64	0.84	0.47	12
65	0.69	-0.16	12
66	0.86	-1.24	15
67	0.70	-1.04	16
68	0.63	-0.69	12
69	0.48	0.27	13
71	0.65	-0.38	13
72	0.54	-1.60	12
73	0.41	-0.38	16
74	0.46	0.69	16
75	0.49	-0.24	16
76	0.26	1.77	14
77	0.71	-1.02	13
78	0.65	-0.26	12
79	0.69	-0.79	14
80	0.40	0.60	12
81	0.60	0.05	13
82	0.80	-0.66	12
83	0.53	-0.19	13
84	0.73	-1.12	12
85	0.39	-0.16	15
86	0.45	-0.36	11
87	0.43	0.74	16
88	0.55	-0.85	11
90	0.68	-0.20	12
91	0.44	0.16	12
92	0.75	-0.56	15
93	0.75	-0.35	12
94	0.38	-0.30	15
95	0.54	-0.40	12
96	0.77	-0.10	16
97	1.05	-0.25	16
98	0.49	1.80	15
99	0.65	-0.07	16
100	0.77	0.52	16
101	0.26	0.59	13
102	0.73	0.16	16
103	0.62	-0.89	16
104	0.97	-0.98	16
105	0.79	-0.10	16
106	0.43	-1.47	15
107	0.50	-0.03	16
108	0.63	0.26	15
109	0.58	-0.46	15
110	0.48	0.18	14
111	0.39	1.89	14
112	0.43	0.43	13

113	0.78	-0.20	12
114	0.52	-1.78	14
115	0.51	-0.18	12
116	0.41	-1.07	13
117	0.65	0.71	13
118	0.51	0.33	13
119	0.81	-1.38	11
120	0.51	-0.94	12
121	0.53	-2.82	13
122	0.46	-1.36	13
123	0.84	-0.91	12
124	0.62	-1.03	12
125	0.70	-1.20	15
126	0.26	-3.20	11
127	0.34	-1.16	13
128	0.46	-0.74	12
129	0.54	-0.98	12
131	0.48	1.53	12
132	0.62	-0.24	12
133	0.35	-1.20	11
134	0.53	-2.00	12
135	0.66	-0.85	14
136	0.81	-1.19	16
137	0.77	-0.37	16
138	0.90	-0.48	16
139	0.42	-0.77	11
140	0.38	1.37	12
141	0.72	-0.71	16
142	0.60	0.16	16
144	0.94	0.35	12
145	0.46	-1.29	11
146	0.53	-0.04	12
147	0.22	0.53	12
148	0.38	0.13	11
149	0.33	-0.35	16
150	0.53	-1.47	12
151	0.56	-0.85	12
152	0.54	0.15	12
153	0.62	-0.32	15
154	0.38	-0.79	15
155	0.41	0.05	13
156	0.61	0.53	13
157	0.37	0.38	14
159	0.30	-0.88	14
161	0.31	-1.44	14
162	0.47	1.55	12
163	0.40	-1.10	14
164	0.49	-0.26	12
165	0.43	-0.67	16
166	0.57	-1.16	11
167	0.36	1.53	11
169	0.38	-0.06	11
170	0.74	-0.71	12
171	0.28	-2.06	12

172	0.29	-0.32	14
173	0.33	-0.17	12
174	0.37	2.35	12
175	0.43	-0.57	14
176	0.58	0.51	12
178	0.57	0.46	12
179	0.23	1.54	11
180	0.30	-2.06	14
181	0.43	0.20	12
182	0.35	-0.59	11
183	0.66	0.23	11
184	0.51	-0.69	12
185	0.76	-0.41	16
186	0.74	-0.28	16
187	0.68	0.62	16
188	0.54	0.52	12
189	0.73	-1.29	11
190	0.55	0.07	14
191	0.51	-0.75	14
193	0.45	0.34	12
194	0.42	-0.53	15
195	0.73	0.37	13
196	0.34	0.08	13
197	0.33	-0.22	13
198	0.24	0.22	13
199	0.42	0.84	13
200	0.65	-1.69	12
201	0.63	-0.75	12
202	0.58	-0.87	14
203	0.34	-1.49	12
204	0.56	-0.17	14
205	0.67	-1.53	12
206	0.69	-0.68	14
207	0.49	-1.08	14
208	0.50	-0.28	12
209	0.44	-0.63	13
210	0.35	0.13	14
211	0.65	-0.06	12
212	0.80	-0.45	12
213	0.39	-0.81	13
214	0.72	-0.71	15
215	0.64	-0.64	16
217	0.48	-0.08	16
218	0.41	-1.59	12
219	0.86	0.45	12
220	0.51	0.84	13
221	0.50	-0.56	15
224	0.35	1.40	14
225	0.32	-0.72	12
226	0.69	0.09	14
227	0.41	0.21	12
228	0.51	0.10	14
229	0.61	0.13	15
231	0.43	0.13	13

232	0.33	2.69	13
233	0.29	-0.74	13
234	0.28	0.21	16
235	0.67	-0.77	16
236	1.03	-0.15	16
237	0.53	-0.52	15
238	0.64	0.47	16
239	1.01	-0.93	16
240	0.79	0.80	12
241	0.52	-0.34	12
242	0.73	-0.16	16
243	0.62	-0.12	16
244	0.28	0.92	14
246	0.47	-0.24	13
247	0.33	-2.11	13
248	0.33	-1.08	12
249	0.43	-0.05	11
250	0.53	-1.96	15

Content Subdomains: A - 11
B - 12
C - 13
D - 14
E - 15
F - 16

BIBLIOGRAPHY

- Adema, J.J. (1990). The construction of customized two-stage tests. Journal of Educational Measurement, 27, 241-253.
- Allen, M.J. & Yen, W.M. (1979). Introduction to Measurement Theory. Monterey, CA: Brooks/Cole.
- Assessment Systems Corporation (1987). MicroCAT testing system. St Paul, MN: Author.
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord & M.R. Novick, Statistical Theories of Mental Test Scores. Reading MA: Addison-Wesley.
- Bock, R.D., Muraki, E. & Pfeiffenberger, W (1988). Item pool maintenance in the presence of item parameter drift. Journal of Educational Measurement, 25, 275-285.
- Boekkooi-Timminga, E. (1987). Some methods for simultaneous test construction. In Wim van der Linden (Ed.), IRT-based Test Construction. (Research Report 87-2). Enschede, The Netherlands: University of Twente, Department of Education.
- Boekkooi-Timminga, E. (1992). Models for Computerized Test Construction. De Lier, The Netherlands: Academisch Boeken Centrum.
- Computing Resource Center (1992). STATA: Statistics/graphics/data management. Santa Monica, CA: Author.
- Computing Resource Center (1989). STAGE: Graphics editor. Santa Monica, CA: Author.
- de Gruijter, D.N.M. (1990). Test construction by means of linear programming. Applied Psychological Measurement, 14, 175-181.
- Fraser, C. (1983). NOHARMII. A FORTRAN Program for Fitting Unidimensional and Multidimensional Normal Ogive Models of Latent Trait Theory. Armidale, Australia: The University of New England, Center for Behavioral Studies.
- Goldstein, H. (1983). Measuring changes in educational attainment over time: Problems and possibilities. Journal of Educational Measurement, 20, 369-377.

- Green, B.F., Bock, R.D., Humphreys, L.G. & Reckase, M.D. (1984). Technical guidelines for assessing computerized adaptive tests. Journal of Educational Measurement, 21, 347-360.
- Gulliksen, H. (1950). Theory of Mental Tests. New York: Wiley.
- Hambleton, R.K. & Cook, L.L. (1977). Latent trait models and their use in the analysis of educational test data. Journal of Educational Measurement, 14, 75-96.
- Hambleton, R.K. & Cook, L.L. (1983). The robustness of item response models and effects of test length and sample size on the precision of ability estimates. In D. Weiss (Ed.), New Horizons in Testing. New York: Academic Press.
- Hambleton, R.K. & Swaminathan, H. (1985). Item Response Theory: Principles and Applications. Boston: Kluwer Nijhoff.
- Hambleton, R.K., Arrasmith, D. and Smith, I.L. (1987). Optimal item selection with credentialing examination. Paper presented at the annual meeting of the American Educational Research Association, Washington.
- Hambleton, R.K., Dirir, M., & Lam, P. (1992). Effects of optimal test designs on measurement precision and decision accuracy. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.
- Harrison, D. (1986). Robustness of IRT parameter estimation to violations of the unidimensional assumption. Journal of Educational Statistics, 11, 91-115.
- Hulin, C.L., Drasgow, F., & Parsons, C.K. (1983). Item Response Theory: Application to Psychological Measurement. Homewood, IL: Irwin.
- Kingsbury, G.G., & Zara, A.R. (1989). Procedures for selecting items for computerized adaptive tests. Applied Measurement in Education, 2, 359-375.
- Kingsbury, G.G. (1990). Adapting adaptive testing with the MicroCAT testing system. Educational Measurement Issues and Practice, 9, 3-6.
- Kingsbury, G.G., & Zara, A.R. (1991). A comparison of procedures for content sensitive item selection in computerized adaptive tests. Applied Measurement in Measurement, 4, 241-261.

- Lord, F.M. & Novick, M.R. (1968). Statistical theories of Mental Test Scores. Reading, Mass: Addison-Wesley.
- Lord, F.M. (1970). Some test theory for tailored testing. In W.H. Holtzman (Ed.), Computer-assisted Instruction, Testing, and Guidance. New York: Harper and Row.
- Lord, F.M. (1980). Applications of Item Response Theory to Practical Testing Problems. Hillsdale, NJ: Erlbaum.
- McBride, J.R. (1976). Bandwidth, fidelity and adaptive tests. In T.J. McConnell, Jr. (Ed.), CAT/C21975: The Second Conference on Computer-assisted Test Construction. Atlanta, GA: Atlanta Public Schools.
- McDonald, R.P. (1980). The dimensionality of tests and items. British Journal of Mathematical and Statistical Psychology, 34, 100-117.
- McDonald, R.P. (1982). Linear versus nonlinear models in item response theory. Applied Psychological Measurement, 6, 379-396.
- Mislevy, R.J. (1986). Recent developments in the factor analysis of categorical variables. Journal of Educational Statistics, 11, 3-31.
- Mislevy, R & Bock, R.D. (1989). BILOG 3: Item analysis and test scoring with binary logistic models. Mooresville, IN: Scientific Software Inc.
- Muraki, E. & Bock, R.D. (1987). BIMAIN: A program for item pool maintenance in the presence of item parameter drift and item bias. Mooresville, IN: Scientific Software.
- Nandakumar, R. (1991). Assessing dimensionality of a set of items - comparison of different approaches. Paper presented at the annual meeting of Division D, American Educational Research Association, Chicago.
- Reckase, M.D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. Journal of Educational Statistics, 4, 207-230.
- Samejima, F. (1973). A comment on Birnbaum's three-parameter logistic model in the latent trait theory. Psychometrika, 38, 221-233.
- Samejima, F (1977). A use of information function in tailored testing. Applied Psychological Measurement, 1, 233-247.

- Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. Psychometrika, 52, 589-617.
- Sykes, R.C. & Fitzpatrick, A.R. (1992). The stability of IRT b values. Journal of Educational Measurement, 29, 201-211.
- Theunissen, T.J.J.M. (1985). Binary programming and test design. Psychometrika, 50, 411-420.
- Theunissen, T.J.J.M. (1986). Some applications of optimization algorithms in test design and adaptive testing. Applied Psychological Measurement, 10, 381-389.
- Thissen, D. & Mislevy, R.J. (1990). Testing algorithms. In H. Wainer (Ed.) Computerized Adaptive Testing: A Primer. Hillsdale, NJ: Erlbaum.
- Urry, V.W. (1977). Tailored testing: a successful application of latent trait theory. Journal of Educational Measurement, 14, 182-196.
- Urry, V.W. (1981). Tailored Testing, its Theory and Practice. Part II: Ability and Item Parameter Estimation, Multiple Ability Application, and Allied Procedures. (NPRDC TR81). San Diego, CA: Navy personnel Research and Development Center.
- van der Linden, W.J. (1987). Automatic test construction using minimax programming. In Wim J. van der Linden (Ed.), IRT-based Test Construction. (Research Report No. 87-2). Enschede, The Netherlands: University of Twente.
- van der Linden, W.J. & Boekkooi-Timminga, E. (1989). A maximin model for test design with practical constraints. Psychometrika, 54, 237-247.
- Verschoor, A. (1991). Optimal Test Design. Arnhem: CITO.
- Wainer, H. & Kiely, G.L. (1987). Item clusters and computerized adaptive testing: a case for testlets. Journal of Educational Measurement, 24, 185-201.
- Wainer, H. & Mislevy, R.J. (1990). Item response theory, item calibration and proficiency estimation. In H. Wainer (Ed.) Computerized Adaptive Testing: A Primer. Hillsdale, NJ: Erlbaum.

- Weiss, D.J. (1982). Improving measurement quality and efficiency with adaptive testing. Applied Psychological Measurement, 6, 473-492.
- Weiss, D.J. & Kingsbury, G.G. (1984). Application of computerized adaptive testing to educational problems. Journal of Educational Measurement, 21, 361-375.
- Weiss, D.J. (1985). Adaptive testing by computer. Journal of Consulting and Clinical Psychology, 53, 774-789.
- Whitely, S.E. & Dawis, R.V. (1976). The influence of test context on item difficulty. Educational and Psychological Measurement, 36, 329-337.
- Wilson, D.T., Wood, R. & Gibbons, R. (1984). TESTFACT: Test Scoring, Item Statistics, and Item Factor Analysis. Mooresville, IN: Scientific Software.
- Wise, S.L. & Plake, B.S. (1989). Research on the effects of administering tests via computers. Educational Measurement, Issues and Practice, 8, 5-10.
- Wise, S.L. & Plake, B.S. (1990). Computer-based testing in higher education. Measurement and Evaluation in Counselling and Development, 23, 3-10.
- Yen, W.M. (1980). The extent, causes, and importance of context effects on item parameters for two latent trait models. Journal of Educational Measurement, 17, 297-311.
- Yen, W.M. (1983). Use of the three-parameter model in the development of a standardized achievement test. In R.K. Hambleton (Ed.), Applications of Item Response Theory. Vancouver, B.C.: Educational Research Institute of British Columbia.
- Yen, W.M. (1985). Increasing item complexity: A possible cause of scale shrinkage for unidimensional item response theory. Psychometrika, 50, 399-410.

