# A FAIR COMPARISON OF THE PERFORMANCE OF COMPUTERIZED ADAPTIVE TESTING AND MULTISTAGE ADAPTIVE TESTING

By

Keyin Wang

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Measurement and Quantitative Methods – Doctor of Philosophy

2017

## ABSTRACT

A FAIR COMPARISON OF THE PERFORMANCE OF COMPUTERIZED ADAPTIVE TESTING AND MULTISTAGE ADAPTIVE TESTING

By

Keyin Wang

The comparison of item-level computerized adaptive testing (CAT) and multistage adaptive testing (MST) has been researched extensively (e.g., Kim & Plake, 1993; Luecht et al., 1996; Patsula, 1999; Jodoin, 2003; Hambleton & Xing, 2006; Keng, 2008; Zheng, 2012). Various CAT and MST designs have been investigated and compared under the same item pool. However, the characteristics of an item pool designed specifically for CAT are different from the characteristics of an item pool designed for MST. If CAT and MST are compared under the same item pool designed for either CAT or MST, the comparison might be unfair to the other test mode. To address this issue, this study focused on comparing the measurement accuracy and averaged test length of MST and CAT, when they were matched on conditional standard error of measurement, exposure rates, IRT scoring method and content specifications, under different item pools designed for MST and CAT, respectively.

When designing a MST, multiple factors need to be considered. In this paper, a total of 16 conditions of MST designs (i.e., 1-2-3 and 1-3-3 panel designs; the AMI and DPI routing strategies; the test lengths of 45 and 60 items; forward and backward assembly) were employed. Each condition was compared with the result of the corresponding CAT. A simulation study was conducted to evaluate the performance of MST against the corresponding CAT.

The results show similar measurement accuracy between MST and CAT, which implies that the efforts to make a fair comparison where successful. The reason is that both procedures matched similar conditional test information. This fair comparison of MST and CAT provides a

reference for testing mode change from CAT to MST in terms of ability recovery and averaged test length. When considering the testing model change from CAT to MST, the backward assembled MST is not suggested even for a classification-oriented test. Whether to change the testing mode depends on the current averaged test length in CAT. If the current CAT has a moderate-length test, switching to a forward assembled MST with 3 stages is plausible and feasible. For a long test, staying in CAT is preferred over switching to MST.

Copyright by
KEYIN WANG
2017

# ACKNOWLEDGEMENTS

First, I want to express extreme gratitude to my amazing advisor and dissertation chair, Dr. Mark Reckase. I thank him for giving me opportunities to work with him on the projects and for guiding me step-by-step. When I transferred from statistics to measurement, I knew little about IRT and all measurement issues. His patience, advice and support did help me grow up to maturity and find my research interest. His profound knowledge and enthusiasm in research really encouraged me to be a good scholar. The way he worked with students and colleagues really motivated me to follow the way he did.

Second, I'm grateful to my advisor in CSTAT, Dr. Steve Pierce, for giving me wise advices in both statistical issues and soft skills. That really helped and prepared me for working outside school. The way he works gave me a good example for how to be a good manager. I wish I could have worked in CSTAT earlier than I did. I also thank him for supporting me when I was doing job-hunting. I truly enjoyed working with my colleagues and advisor in CSTAT.

Next, I would also like to thank the rest of my dissertation committee members, Dr. Kim Maier, Dr. Spyros Konstantopoulos and Dr. Yuehua Cui for helping me throughout the entire process. I thank Dr. Spyros Konstantopoulos for giving me an offer to MQM. I also thank Dr. Yuehua Cui for giving me advices for the coming Ph.D. career when I was graduating with a Master's Degree, and for supporting me when I looked for a summer internship.

I had worked in the K-12 Outreach for two and half years. I would like to express my sincere appreciation to Dr. Neelam Kher, Jacqueline Gardner and Kathleen Wight. I would attribute most of my knowledge about the K-12 education to their generous help and guidance. I truly enjoyed working with them and we have become good friends.

I would also like to thank my colleagues in MQM: Xin Luo, Liyang Mao, Chi Chang, Tingqiao Chen, Jiahui Zhang, Emre Gonulates, Wei Li, Xuechun Zhou, Unhee Ju for supporting me in both life and study. I couldn't enjoy my Ph.D. career without them in my life.

I have been blessed with a wonderful group of people while I was doing a summer internship in Measured Progress. I thank each person in the psychometric team there. They set a good example for me of how to work with colleagues. I especially thank Dr. Wonsuk Kim and Dr. Louis Roussos for their generous help and support.

Now, I would like to thank my church family at Lansing Chinese Christian Church (LCCC). My dear group of brothers and sisters at LCCC has been my extended family for the past eight years. I cannot imagine what my life would be without my brothers and sisters at LCCC. I could not overcome the difficulties in both life and study without their love, help and prayer. I especially thank the brothers and sisters who helped me in running the analyses of my dissertation. They are Yingqian Lin, Xiaoge Wang, Shiyao Liu, Tim Lin, Shutian Yan, Oscar Xu and Yuelin Wu. Without their generous help, I could not defend my dissertation in time.

Throughout my whole life, I'm blessed with loving family members. I would like to dedicate my dissertation to my family:

To my grandparents, whom had taken care of me for years when I was little. They love me but don't spoil me. They are the persons who helped me set a good study attitude and encouraged me to work hard through every oversea phone call.

To my father Jinhe Wang, and my mother Ping Wang, who spent thirty years raising me, who have been there to support me financially and emotionally, who love me unconditionally and who inspired me to pursue my own life abroad.

To my aunt Li Wang, who supports me a lot since I have been in the US. Thank you to all my families! This accomplishment is as much mine as it is yours.

Last but not the least, I want to give many thanks to the Lord Jesus Christ for Your saving grace and mercy, despite my frequent lack of faith. I was led by God to my current major which is a miracle to me. I did not know what MQM represents and what measurement is when I was graduating from the Department of Statistics with a Master's Degree. But God prepared my current major and my entire Ph.D. career for me, and guided me though the past six years. I pray that I will follow the path you have set for me and will live for you in my whole life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

# KEY TO ABBREVIATIONS

3PLM: Three-parameter Logistic Model

AICPA: The American Institute of CPAs

AMI: Approximate Maximum Information

ATA: Automated Test Assembly

CAT: Computerized Adaptive Testing

CCAT: Constrained Computerized Adaptive Testing

CPA: Certificated Public Accountants

DPI: Defined Population Intervals

EAP: Expected a Posteriori

GAMT: Graduate Management Admission Test

GRE: Graduate Record Examination

IRT: Item Response Theory

LFT: Linear Fixed Length Test

LSAT: Law School Admissions Test

MAP: Mode of the Posterior

MCCAT: Modified Constrained Computerized Adaptive Testing

MI: Maximum Information

MIP: Mixed Integer Programming

MLE: Maximum Likelihood Estimation

MPI: Maximum Priority Index

MSE: Mean Squared Error

MST: Multistage Adaptive Testing

NAEP: National Assessment of Educational Progress

NC: Number-Correct

NCLEX: National Council Licensure Examination

NWADH: Normalized Weighted Absolute Deviation Heuristics

SH: Sympson-Hetter

STA: Shadow-Test Approach

TIF: Target Test Information Function

USMLE: U.S. Medical Licensure Examination

WDM: Weighted Deviation Model

# CHAPTER 1 INTRODUCTION

## 1.1 Background

Since the early 1970s, computerized adaptive testing (CAT) has been extensively researched and implemented in educational assessments. Starting in the 1990s, the Graduate Management Admission Test (GMAT) and the National Council Licensure Examination (NCLEX) have changed successfully from paper-and-pencil (P&P) format to CAT format (Gu, 2007). In operational CATs, each examinee is administered a tailored test with the items well matching their estimated ability level. After administering each item within the test, an examinees' estimated ability level will be updated for selecting the next item. The process of administering items does not stop until a certain measurement accuracy is achieved or until the maximum test length is reached. The main advantage of CAT over paper-and-pencil tests is achieving measurement precision with shorter tests. A shorter test can reduce examinees' fatigue that may have an impact on their test results. In addition, the computers delivering CAT are able to give immediate scoring feedback, to have flexible testing schedules, and to adopt new item formats (Chalhoub-Deville & Deville, 1999). Although there are multiple advantages of CAT, the disadvantages of CAT have aroused researchers' concerns. First, the test form will not be assembled until the end of the test. It is impossible for test specialists to review each test form for test quality purposes. The quality of item pool can be guaranteed, but the one of an individual test form cannot. Second, the examinees are not allowed to review their answers on previous items, which is the greatest disadvantage of CAT (Lunz, Bergstrom & Wright, 1992).

To eliminate the disadvantages of CAT, multistage adaptive testing (MST) as an alternative has been increasingly developed and implemented. The Certificated Public Accountants (CPA) Examination, the Graduate Record Examination (GRE), the Law School Admissions Test

(LSAT), the U.S. Medical Licensure Examination (USMLE) and the National Assessment of Educational Progress (NAEP) have switched successfully from paper-and-pencil (P&P) and CAT formats to MST formats. Although there are various terms applied to MSTs including multistage testing (Patsula, 1999), multistage adaptive testing (Zheng et al., 2012), and computerized multistage testing (Ariel, Veldkamp & Breithaupt, 2006), multistage adaptive testing (MST) will be used in this study. Before MST administration, multiple panels are developed for test security purposes. Groups of items known as test modules are preassembled in each panel with several stages (Luecht & Nungster, 1998). In the beginning of MST administration, each examinee receives a randomly selected test panel. In each panel, examinees receive a module at each stage of testing. They are assigned to the next module with pre-determined routing rules according to their performance on the previous stage. This is where the point of "adaptive" derives from. MST involves adaptive selection of a group of items instead of adapting every item. The number of stages and modules are the same across panels, but both numbers in each stage can vary in different test designs.

There are several advantages of MST compared to CAT. First, as mentioned by Wainer (1990), the distinct advantage of MST over CAT is that test developers are able to review a small number of pre-assembled MST test forms before delivery for quality control, rather than simply relying on the adaptive algorithm to form the test. Second, MST maintains a lot of the advantages of CAT such as providing information on speed of response, convenient test scheduling with individuals, and immediate scoring feedback. Third, examinees are allowed to review and change their answers to the previous items within modules in a MST administration. Scoring and routing procedures are implemented after examinees submit their module.

## 1.2 Statement of Problem

From the review of recent literature comparing the performances of CAT and MST, various CAT and MST designs are investigated and compared under the same item pool. The study of Kim and Plake (1993), Patsula (1999), Armstrong et al. (2004) and Keng (2008) consistently indicated that the CAT design is more accurate and efficient than the MST design. However, the characteristics of an item pool designed specifically for CAT are different from the characteristics of an item pool designed specifically for MST. A CAT item pool should contain an appropriate number of items to build individualized tests according to each examinee's ability level. A MST item pool should have sufficient items to meet the specification of the automated test assembly process and reflect the demands of measurement accuracy. If CAT and MST are compared under an item pool designed for CAT, the comparison might be unfair because the item pool is in favor of CAT. Similarly, if the comparison is conducted under an item pool designed for MST, the item pool could also be in favor of MST. To address this issue, this study assembled one item pool for MST and the other for CAT from the master pool based on their test design. Thus, the primary purpose of this study is to compare the performance of MST and CAT with matching psychometric properties under two separate item pools constructed from a master pool.

While administering a MST, the measurement precision can be affected by a number of factors including the number of stages, the number of modules per stage, module length, and the distribution of item difficulty per module (Zenisky & Hambleton, 2014). Longer test lengths tend to give a more accurate ability estimate. Specifically, if the routing module is not long enough to produce an accurate trait level, examinees could be routed to a wrong next module. Routing Strategy helps to determine the routing points for assigning examinees to next modules. The

assembly priority is a factor that cannot be ignored during the test assembly process. When selecting items to fit the pre-determined test information function from the available items in a pool, modules assembled earlier are likely to have a better fit than modules assembled later. Various MST designs were always conducted and evaluated before implementing a final design in practice. Thus, another purpose of this study is to investigate and compare different MST designs in terms of measurement precision. Each MST design will be compared with CAT.

Therefore, the research questions of this study are:

1) Does MST outperform CAT consistently in terms of measurement precision, when matching similar test information on overall ability scale, item exposure rate and test content specification?
2) Which MST designs will give the highest measurement precision under different conditions (e.g. two levels of panel designs, two levels of assembly priority, two levels of routing strategy and two levels of test lengths)?
3) Which testing mode (i.e., CAT or MST) will give a shorter test under each item pool with matched properties? Will the result of comparison be consistent across all conditions?

# CHAPTER 2 LITERATURE REVIEW

This chapter contains three sections. First, a brief description of the procedure for computerized adaptive testing (CAT) administration and main factors considered are presented. Second, a brief description of the procedure for multistage adaptive testing (MST) is introduced, and followed by the design considerations of a MST. Related to the current topic, this study focuses on the explanation of item pool, panel consideration design, test assembly and routing strategy. The third section provides a review of several current comparison studies between CAT and MST.

## 2.1 Computerized Adaptive Testing (CAT)

Computerized adaptive testing has been widely used in educational testing programs. CAT is a method of administering items sequentially according to the ability level (θ) of each examinee. In CAT, each item is selected by a pre-determined item selection rule according to the examinee's current ability estimate ($\hat{\theta}$), based on the available responses in the test. Then, the difficulty of each item is well matched to the examinee's ability level and other practical requirements such as content balance and item exposure rate. The process of selecting items continues until the stopping rule is met. He (2010) provided a clear presentation of the adaptive nature of CAT, which is displayed by Figure 2.1.

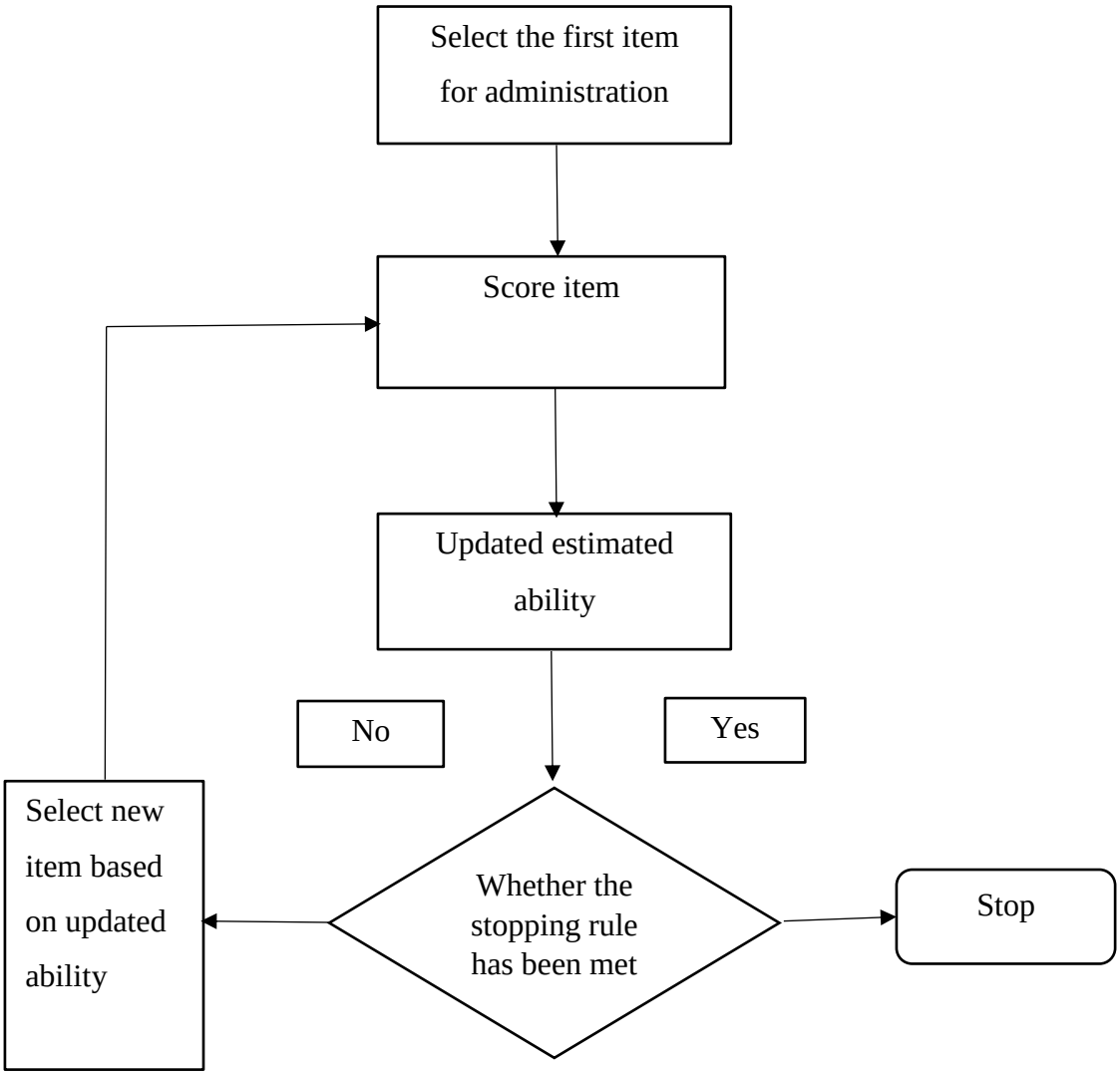


Figure 2.1 Steps for Administrating a CAT (He, 2010)

Reckase (1989) stated four major components of CAT: the item pool, the item selection procedure, the scoring procedure and the stopping rule. In addition, some additional components, such as content balancing and item exposure control, are always incorporated in the item selection procedure. Generally, CAT administers a tailored test by selecting items well matched to examinee's estimated ability level, and then achieve a desirable measurement accuracy with a short test length.

### 2.1.1 CAT Item pool

CAT requires an item pool to have sufficient number of high quality items to give thousands of tests to examinees. In reality, there are two kinds of item pools: the master pool and the operational pool. The master pool has as many items as possible to supply the testing program. The operational pool is used to give individual tests during a test administration period. The range of difficulty of items in an operational item pool should cover the range of examinee's ability levels to ensure that all the examinees receive items well matched to their ability levels (Gu, 2007). In addition, exposure control and content balancing are required to be incorporated in the process of designing an item pool. A large item pool size is always suggested to provide accurate ability estimate over a broad range, to avoid item over-exposure and to maintain content balancing (Patsula & Steffan, 1997; Luecht, 1998; Luo, 2015). However, item writing cost and effort also need to be considered in practice.

### 2.1.2 Item Selection Procedure

The item selection procedure is a basic component of a CAT. The widely used item selection procedures in CAT are the maximum information method (MI; Weiss, 1982) and Bayesian approach (Owen, 1975). The MI selects the next item providing the maximum Fisher information on the current ability level. For a given dichotomous item *j*, Fisher information is (Lord, 1980):

$$I_j(\theta) = \frac{\left[\frac{\partial P_j(\theta)}{\partial \theta}\right]^2}{P_j(\theta)(1-P_j(\theta))} = \frac{\left[P_j'(\theta)\right]^2}{P_j(\theta)(1-P_j(\theta))} \quad (2.1)$$

where $P_j(\theta)$ denotes the probability of correct response on item *j* given θ. In the case of the unidimensional three-parameter logistic model (3PLM), the Fisher Information for a dichotomous item is (Lord, 1980; Hambleton, & Swaminathan, & Rogers, 1991):

$$I_j(\theta) = \frac{D^2 a_j^2 (1-c_j)}{(c_j + e^{Da_j(\theta - b_j)})(1 + e^{-Da_j(\theta - b_j)})^2} \quad (2.2)$$

where $D$ =1.7, $a_j$ is the item discrimination parameter, $b_j$ is the item difficulty parameter, and $c_j$ is the pseudo-guessing parameter. This item selection method prefers the item with a large discrimination parameter because this provides large item information at the current ability level.

The Bayesian approach is to select an item that maximizes the expected posterior precision of the ability estimate. Chang & Stout (1993) presented that the Bayesian approach may select different items then the MI in the early stage of CAT, but gave similar result with the MI as the test length increases. Various research compared the performances of item selection methods, and found no difference between the MI and the other method (Veldkamp, 2003; Ho, 2010). Thus, MI is adopted in this study for convenience.

2.1.3 Scoring Procedure

One of the advantages of CAT is selecting the item well matched to an examinee's ability level. In the beginning of the test, an initial value of the ability level is arbitrarily set since there is no information about an examinee. Then, the ability estimate is updated repeatedly after administering each item based on the available responses at that time. The two widely used ability estimation methods are maximum likelihood estimation (MLE) and Bayesian estimation. The MLE method finds the ability estimate according to the maximum value of the likelihood function,

$$L(\boldsymbol{u}|\theta) = \prod_{i=1}^{n} P_i(u_i|\theta, a_i, b_i, c_i) \quad (2.3)$$

where $n$ is the number of items, and $P_i(u_i|\theta, a_i, b_i, c_i)$ is the probability of getting response $u_i$ ($u_i$=0 for incorrect response and 1 for correct response) on item $i$ given item parameters and

examinees' true abilities. The MLE method provides a $\hat{\theta}$ as the estimate of examinee's true ability by setting the first derivative of $L(\boldsymbol{u}|\theta)$ as 0,

$$\frac{\partial}{\partial\theta} L(\boldsymbol{u}|\theta) = 0 \tag{2.4}$$

However, this method would give an infinite ability estimate if the item responses are all correct or incorrect at the early stage of CAT. In reality, the arbitrary minimum and maximum ability estimates (e.g., -4 and +4) for such response patterns are set to solve this problem. MLE cannot be used until one correct or one incorrect response are obtained. Bayesian estimation is also considered as an alternative to MLE for solving the infinity problem, because it can estimate examinees' ability after the first response. In the Bayesian estimation procedure, the posterior distribution of ability level is updated based on Bayes Theorem with the specified prior distribution

$$f(\theta|\boldsymbol{u}) = \frac{f(\boldsymbol{u}|\theta)f(\theta)}{f(u)} \tag{2.5}$$

where $f(\theta|\boldsymbol{u})$ is the posterior distribution, $f(\theta)$ is the prior distribution, and $f(\boldsymbol{u})$ is the likelihood of a given response string *u* which is a constant. The mean of the posterior distribution (EAP) or the mode of the posterior distribution (MAP) is used to update the ability estimate. Although the Bayesian method can solve the problem of MLE, one disadvantage is that the selection of prior distribution may have an impact on the final ability estimate. Wang & Vispoel (1998) pointed out that if an inappropriate prior is selected, the final estimate could be biased a lot. Various studies by (Chen, Hou, & Dodd, 1998; Wang & Wang, 2001; Ho, 2010) have compared different ability estimation methods and conclude that the MLE has comparable effect

on the results with other methods. Therefore, this study applied MLE to estimate the examinee's ability level.

### 2.1.4 Content Balancing

The procedure for meeting the constraints of content area and item format is called the content balancing procedure. Taking the adaptive test, the test-takers must receive the same distribution of items by content area to obtain relative comparable test scores (Stocking & Swanson, 1993). In operational CATs, the content balancing procedure is always implemented through the item selection algorithm (Kingsbury & Zara, 1991). Various approaches have been researched and applied in CAT, such as the constrained CAT approach (CCAT; Kingsbury & Zara, 1991), the weighted deviation model approach (WDM; Swanson & Stocking, 1993), the shadow-test approach (STA; Van der Linden & Reese, 1998), the modified CCAT (MCCAT; Leung, Chang & Hau, 2003b) and the maximum priority index approach (MPI; Cheng & Chang, 2009). Generally speaking, the STA, the WDM and the MPI are more flexible in dealing with a number of constraints (He, 2010).

### 2.1.5 Exposure Control

When administering items in an adaptive test, examinees with similar abilities tend to receive multiple overlapped items. Additionally, selecting the most informative items frequently will possibly allow examinees to remember some items and circulate them to future examinees. Then, the future examinees will collect this pre-knowledge, which reduces the precision of measurement. To prevent the leaking of the information to the future examinees, it is therefore important that the exposure rate of each item should be controlled below a threshold. Way (1998), Davis & Dodd (2003) and Davis (2004) classified exposure control procedures into three categories: randomization, conditional selection and stratification procedures.

The randomization procedure randomly selects the next item from a group of items near optimal level instead of selecting the item with maximized information at the ability level. The 5-4-3-2-1 technique (McBride & Martin, 1983) and the randomesque method (Kingsbury & Zara, 1989) are two widely used randomization procedures. They are very straightforward to implement, but do not guarantee the exposure rates of items will be constrained to a given level (Davis, 2004; Keng, 2008).

Conditional selection procedures control the exposure rate of an item on a given criterion which is the exposure control parameter. This parameter guarantees the maximum exposure rate. The Sympson-Hetter (SH) method (Sympson & Hetter, 1985) is the most commonly used conditional selection procedure. This method assigns an exposure control parameter to limit the maximum exposure rate of each item to a predetermined level. During the test administration, if the exposure control parameter is greater than a random number, the item will be selected. However, one disadvantage of SH method is that implementation is very time-consuming for determining the exposure control parameters (Keng, 2008).

### 2.1.6 Stopping Rule

Fixed length and variable length are two ways to decide when to terminate the test. A fixed length test requires all the examinees to take the same number of items in a test, which takes similar testing time across examinees. However, one disadvantage is that test reliability reporting will have a problem due to different measurement precision across examinees (Gu, 2007). A variable length test has examinees take different numbers of items until a pre-specified precision level of ability estimate is met. A target standard error of measurement can be used as a stopping criterion to terminate the test, so that each examinee can be measured to the same degree of precision. Compared with fixed-length tests, variable-length tests tend to improve the item pool

use due to minimizing test length (Bergstrom & Lunz, 1999). One problem in variable-length test is that the examinees with extreme ability levels will have a longer test than the one with ability levels matching the items in the pool. The item pool will possibly run out of appropriate items to administer. Thissen & Mislevy (2000) suggested that the combination of specific precision and maximum number of items should always be implemented in practice to avoid this situation.

## 2.2 Multistage Adaptive Testing (MST)

The basic components of a MST are similar to those of a CAT, such as an item pool from which all test forms are built, routing strategies assigning examinees to the next module, scoring methods to report an examinee's final score, and test specification to construct the test forms. In addition, MSTs also have some unique factors, such as modules, panels and stages.

As defined earlier, modules are bundles of items that are built before test administration. Each module can be built to meet both a statistical target like test information function or a non-statistical target like content specification. According to the overall difficulty level of items in a module, a module can be classified to easy, moderate and hard categories. Once modules are built, they are combined to create a panel for administration. A panel is analogous to a test form since it needs to meet both statistical and non-statistical targets. Multiple panels should be built for controlling the exposure of modules and items. Each examinee will be assigned one panel in a MST administration. In addition, a series of stages also exist in a panel. Most MSTs have two to four stages. Each stage has a number of modules. The first stage of a MST always has one module taken by each examinee. The later stages can have multiple modules.

Figure 2.2 gives an example of the ten parallel panels having three stages and seven modules (i.e., 1-3-3 design). As its name suggested, the term 1-3-3 means that one module in the first stage, and three modules in stage 2 and 3 within each panel. The letters E, M and H represent the

average difficulty of the module (E = Relatively Easy, M = Moderately Difficulty, H = Relatively Hard). The possible pathways across the modules are identified by solid lines. How to assemble multiple parallel panels was discussed in the next chapter.

Any examinee who is administered a 1-3-3 design MST will take the items in Module 1M first. Based on the performance in Module 1M, examinees will be routed to one module of stage 2. If examinees perform well, then will be routed to the hard module (Module 2H); if examinees perform moderately well, then will be routed to the moderate difficult module (Module 2M); and if examinees perform poorly, then will be routed to the easy module (Module 2E). Luecht & Nungester (1998) suggested that the extreme change of performance from L to H is unlikely to happen, and thus examinees cannot be routed from easy to hard or hard to easy module. Similar rules will be applied to route examinees from stage 2 to 3.

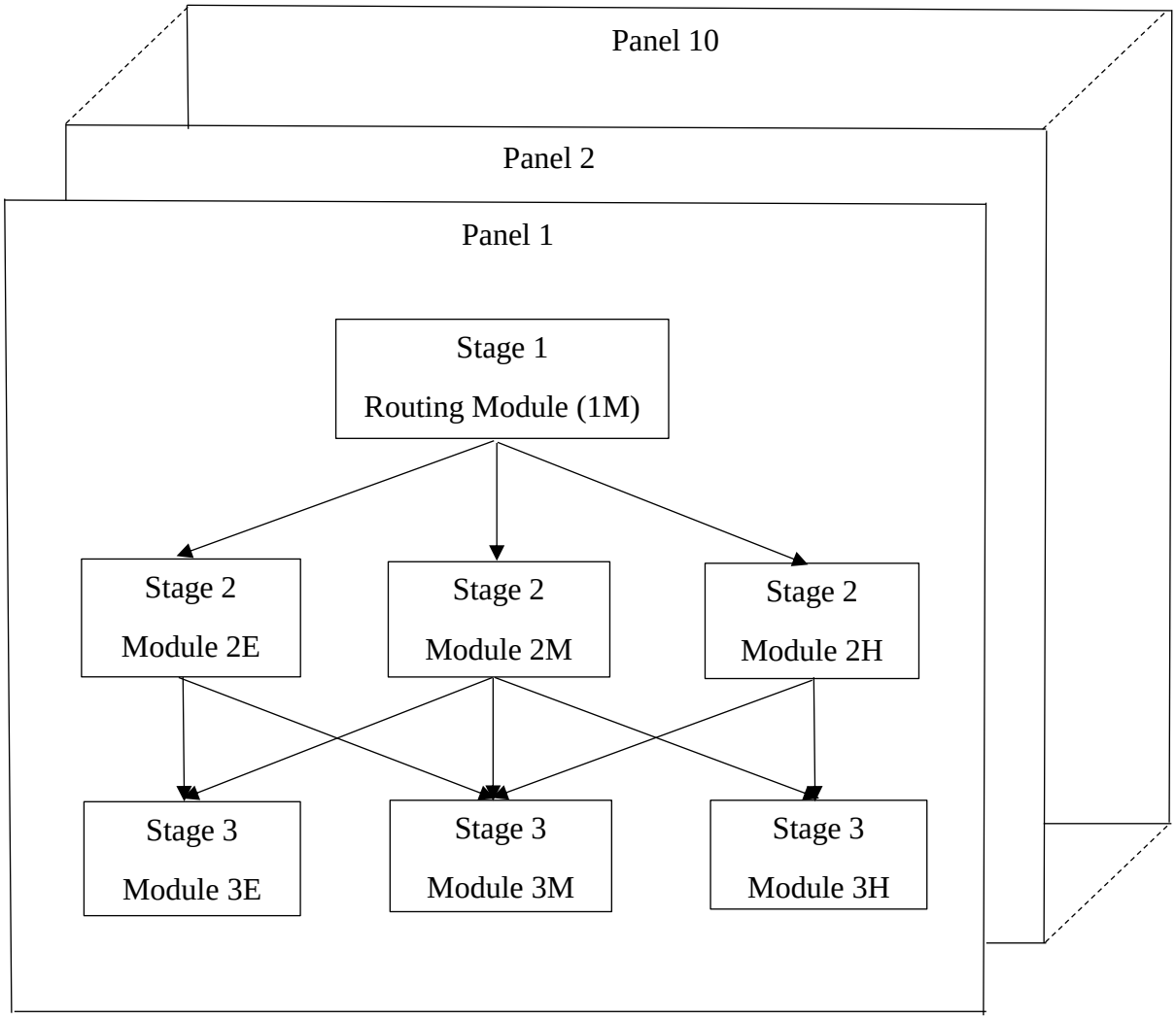


Figure 2.2 Example of 1-3-3 Multistage Test with 10 Panels. The solid lines are the possible pathways for an examinee. E = Relatively Easy; M = Moderately Difficult; H = Relatively Hard.

When designing a MST, a number of considerations about the components should be decided, such as item pool, panel configuration, test assembly, routing strategy and scoring procedure. In addition, content specification and examinees' ability distribution are incorporated in making these decisions. Details of these considerations are explained as followed.

### 2.2.1 MST Item pool

The item pool, which is built to incorporate the test content and both statistical and non-statistical constraints, is considered an important factor to obtain measurement results in MST. Hendrickson (2007) suggested that the item pool should have a sufficiently large size to support assembling modules and multiple panels. Luecht and Nungester (1998) suggested that the quality of item pool affects the successfulness of the test assembly process. Wang (2012) indicated that using an item pool specifically designed for MST contributes in scoring accuracy. For example, two studies have examined the effect of item pool characteristics on psychometric properties of MSTs. Jodoin (2003) compared three computerized test designs under two item pool conditions: 1) item pool quality measured by item discrimination, which is specified as low-, middle- and high-discriminating item pool; 2) the match between test content specification and item pool. The results showed that measurement precision and classification accuracy benefit from the item pool quality and the match between test specification and the item pool. Wang et al. (2012) compared 25 different panels under both an item pool designed for fixed form tests and an "optimal" item pool for MST. The results indicated that the quality of item pool affects the quality of panel design. They also suggested that different panel configurations need different optimal item pools for that design.

Generally speaking, a MST item pool supports test assembly as a CAT pool does. The ideal item pool should have sufficiently large size to well target desired difficulty range and to provide flexibility for module and panel assembly. But item writing cost and item exposure rate are always considered in controlling the item pool size. Wang et al. (2012) suggested that the MST item pool size is determined as 1.5 times of the required number of items in test design.

### 2.2.2 Panel Configuration

A panel is considered as a combination of modules. Panel design configuration can vary in the following ways: 1) the number of stages; 2) the number of module per stage; 3) the length of each module; 4) the distribution of item difficulty among modules; and 5) item-attribute and test requirements for the modules at each stage (Luecht & Burgin, 2003). Generally, design decisions about these issues depend on the factors such as the purpose of the test, available items and test specifications.

Different across-stage module arrangements (e.g. simple 1-2 and 1-3, 1-2-2, 1-2-3, and 1-3-3) are researched in the MST literature. Patsula (1999) suggested that increasing the number of stages from two to three improves measurement precision. Jodoin et al. (2006) showed that a two-stage 40-item MST provided a lower measurement precision and classification accuracy than a three-stage 60-item MST, which reinforced that increasing the number of stages generally increases measurement precision. Hendrickson (2007) also noted that recent studies have been using three or four stages instead of simply two stages.

In terms of the number of modules in each stage, most MST research use one module in the first stage. Lord (1971) and Kim & Plake (1993) found that the number of modules in stage 2 affect measurement accuracy. The results of Amstrong et al.'s paper (2004) indicated that three modules per stage is sufficient for desirable accuracy of ability estimates for most MSTs.

In terms of the number of items within each module, Kim & Plake (1993) found that a longer first-stage module which is also known as the routing test contributes to the accuracy of the ability estimate. Within a fixed-length MST, Patsula (1999) showed that varying number of items within each module did not affect the accuracy of ability estimation.

In MST, the difficulty level of the module is targeted at a specific range on the ability scale (e.g. low, medium and high difficulty level), which is also an implicit part of the assembly process of selecting items to match the desired test information function for each module (Ariel, Veldkamp, & Breithaupt, 2006).

### 2.2.3 Test Assembly

The test assembly in MST focuses on incorporating statistical and non-statistical test specifications simultaneously with an item pool and mathematical algorithm to select items and construct multiple modules and panels (Zenisky & Hambleton, 2014). To achieve this goal, automated test assembly (ATA) is an effective way to assemble modules from the existing item pool (Melican, Breithaupt, & Zhang, 2010).

Luecht and Nungester (1998) proposed two heuristic "Top-Down" and "Bottom-Up" assembly strategies to build panels for MST. The Top-Down strategy (e.g., Zheng et al., 2012) is used to assemble not completely parallel modules which are combined to meet test-level constraints. That is, modules are not exchangeable across panels when the Top-Down strategy is used. The Bottom-Up strategy (e.g., Luecht et al., 2006) is applied to assemble parallel forms of each module, and then mixed and matched these modules to build parallel panels. Each module is built up independently by meeting module-level content requirements and statistical constraints. Thus, modules within the Bottom-Up strategy are exchangeable across panels. Compared to the Top-Down strategy, which deals with uneven constraints for each module, the Bottom-Up assembly is easier and more straightforward to implement.

When constructing a MST, the widely used statistical target is the target test information function (TIF). The target TIF is a pre-determined curve to specify the amount of required test information. Luecht & Burgin (2003) pointed out that target TIFs need to reflect three goals: 1)

to help guarantee measurement precision provided by test information functions; 2) to derive targets that produce large number of content-balanced module; and 3) to control the conditional exposure of items in a test among examinees. Luecht and Nungester (1998) suggested the shape of TIF to be built sharper in the later stage according to the decreasing standard deviation of item difficulty across modules. The reason is that the later modules focus on a narrow range of examinee's ability. The Target TIF is also used to determine the cutoff point on ability scale to route examinees when applying the approximate maximum information (AMI) method. Once the TIFs and other non-statistical constraints are determined, multiple panels can be constructed simultaneously by a test assembly method.

Linear programming and heuristic methods are commonly used to assign items to each module and create panels. The linear programming method is able to strictly satisfy a number of constraints (e.g., content specifications and item exposure rate) when building the panels (Armstrong et al., 2004; Van der Linden, 2005; Breithaupt and Hare, 2007; Luetch et al., 2006). The essential part of this method is to provide optimal solutions of a set of inequalities, which are the assignments of items into modules. Mixed Integer Programming (MIP) is one well-known form of linear programming methods, which can be found in van der Linden (2005), Breithaupt and Hare (2007), and Melican, Breithaupt, and Zhang (2010). The MIP can have a large number of feasible solutions (all constrains are met), and then find the best possible solution. Some MIP based software have been developed, such as CASTISEL (Luecht, 1998), CPLEX 9.1 (ILOG, 2005), JPLEX (Park, Kim, Dodd & Chung, 2011) and lp_Solve (Diao & Van der Linden, 2011a). Linear programming methods for ATA are widely used to conduct MST test assembly, but the detailed discussion of it goes beyond this study.

Heuristic test assembly method generally includes the weighted deviation model (WDM; Swanson and Stocking, 1993), the normalized weighted absolute deviation heuristics (NWADH; Luecht, 1998) and the maximum priority index (MPI; Cheng & Chang, 2009).

The WDM incorporates both statistical and non-statistical item properties by the user-assigned weights to achieve a desirable measurement balance. Deviation from the content targets is weighted with the deviation from the current test information to the target but unreachable value. The WDM selects the item with the smallest sum of weighted deviations for a CAT administration. The item selection algorithm using WDM includes three steps generally. First, the deviation for each of the constraints is calculated by assuming each item was already selected to the test. Second, the weighted deviation across constraints are summed. Finally, the item with the smallest summed deviation will be administered. A comprehensive description of this heuristic method is provided in Swanson & Stocking (1993).

The NWADH is the only heuristic method used in MST assembly studies (e.g. Luecht & Nungester, 1998; Patsula, 1999; Jodoin et al., 2006; Hambleton & Xing, 2006; Wang, 2014). This heuristic has been successfully implemented in a medical licensure test assembly problem. Compared to MIP, it doesn't need commercial software but can always provide a solution to meet all constraints. In addition, the ease and speed of converging to a feasible solution might be more valuable than converging to a best possible solution (Luecht, 1998; Swanson & Stocking, 1993). Thus, this study chose the NWADH to conduct MST test assembly for simplicity and feasibility. The MPI is able to account for both statistical and non-statistical constraints as well, which was applied in this study to assemble individualized tests in CAT. The details of these two methods were discussed in the following section.

### 2.2.4 Routing Strategy

The routing strategy in MST assigns examinees to the next well-matched module on the basis of their performances in the previous module, which is analogous to the items selection procedure in CAT. Two routing strategies have been widely used: the approximate maximum information strategy (AMI; Luecht, Brumfield & Breithaupt, 2006) and the defined population intervals strategy (DPI; Jodoin et al., 2006; Zenisky & Hambleton, 2014). The former one identifies the empirical target TIF first, and then sums the TIFs of a previous administered module and current alternative modules respectively. The next step is to find the intersection point of adjacent cumulative TIFs as the routing point. This method is analogous to the maximum information item selection strategy in CAT, given a current provisional estimate. The later one, which relates to policy issues, routes a pre-specified proportion of examinees to the next modules according to their rank-ordered ability estimates. The value of the proportion is predetermined. For example, in a 1-3-3 panel design, if roughly equal number of examinees are required to be assigned to module 2E, 2M and 2H, the scores of the 33 percentile and the 67 percentile would be the routing points. According to the normally distributed population, those would be -0.44 and 0.44 on the θ-scale.

Zenisky, et al (2010) suggested the commonly used scoring methods included IRT-based proficiency estimate, number-correct (NC) scoring and cumulative weighted NC. IRT-based scoring, which is usually done by maximum likelihood estimation (MLE) or expected a posteriori (EAP) estimation, is commonly used in various research studies (e.g. Kim & Plake, 1993; Jodoin, 2003; Jodoin et al., 2006; Hambleton & Xing, 2006 and Keng, 2008). Even though NC scoring is straightforward to route examinees, it is inappropriate to be reported as the final

ability estimate. The reason is that examinees take statistically nonequivalent items in a MST (Lord, 1980). Thus, this study applied IRT-based scoring for reporting the final ability estimate.

### 2.3 Comparison of CAT and MST

Various recent studies have investigated the comparison of item-level CAT and MST in terms of some psychometric properties. Kim & Plake (1993) compared the measurement precision and relative efficiency between CATs and MSTs with the length of first-stage module (10, 15, 20 items), total test length (40, 45, 50 items), number of second-stage modules (6, 7, 8 modules), and distribution of item difficulty in the first-stage module (peaked or rectangle) varied. The results indicated that CAT outperformed MST in terms of both measurement precision and relative efficiency.

The study by Patsula (1999) was conducted to compare the accuracy of ability estimation in different item-level CAT designs, P&P design, and the MST designs in terms of number of stages, number of modules per stage, and number of items per module. The study noted that item-level CAT produced the most accurate ability estimate over P&P and MST, and that increasing the number of modules per stage increased measurement precision and efficiency of MST.

Jodoin (2003) compared linear fixed length test (LFT), CAT and MST with item pool characteristics, degree of match between test and item pool content specifications, total test length and exposure control varied. The results, not surprisingly, indicated that the CAT design outperformed MST and LFT designs in terms of some psychometric properties.

Hambleton and Xing (2006) compared the performances of computer-based LFTs, CATs and MSTs in terms of classification accuracy, and found once again that CAT performed the best, followed by MST and then LFT.

Keng (2008) investigated the performance of the innovative testlet-based CATs, item-level CATs and MSTs. The results indicated that MST yields good measurement accuracy, good item pool utilization but high item exposure rates. In general, the results in the literature confirmed that CAT results achieve better than those of MSTs in terms of measurement precision.

# CHAPTER 3 METHODOLOGY

This chapter first presents the various conditions for the MST simulation study. Then, the simulation study of CAT was matched to some psychometric properties of MST design. All the procedures were completed using Matlab R2011a Student version, R2015b and R2016a Academic Version.

## 3.1 MST Simulations

Several conditions of MST test designs were compared in this simulation study, including test length, panel design consideration, routing strategy and assembly priority. Within the operational MST pool, sixteen separate combination conditions were generated across ten panels. Each one was compared with the result of the corresponding CAT. Ten replications were implemented to obtain stable estimates. The list of factors that were varied for the simulations of MST and CAT is presented in Table 3.1. The explanations for the selection of the factors are presented below.

Table 3.1 Simulation Factors

| | MST | CAT |
|---|---|---|
| Master pool size | 8100 | 8100 |
| Item pool size | 2700 | 900 |
| Panel design | 1-2-3; 1-3-3 | ---- |
| Test length | 45; 60 | Variable-length |
| Routing strategy | AMI; DPI | MPI |
| Assembly priority | Forward; Backward | ---- |

### 3.1.1 Panel Configuration

As stated earlier, the paper by Amstrong et al. (2004) showed that three modules per stage was sufficient for desirable accuracy of ability estimates for most MSTs. In addition, testing agencies prefer to implement a three-stage MST at minimum for operational tests since it provides a second routing point (Zenisky and Hambleton, 2014). Then, this study applied the 1-2-3 design

(Zenisky 2004; Armstrong & Roussos, 2005) and 1-3-3 design (Zenisky 2004; Luecht et al 2006; Jodoin, Zenisky & Hambleton 2006). The 1-3-3 design is described in the previous chapter. The 1-2-3 design is illustrated in Figure 3.1 below. As mentioned earlier, the letters E, M and H represent the average difficulty of the module (E =Relatively Easy, M = Moderately Difficulty, H = Relatively Hard). The possible pathways across the modules are indicated by solid lines. Three-stage panel designs are not the only commonly-used designs. Because Stark and Chernyshenko (2006) suggested that the greatest influence on measurement precision of MSTs did not lie in most of test design considerations but the test length, this study only adopted three-stage panel design with test length varied.

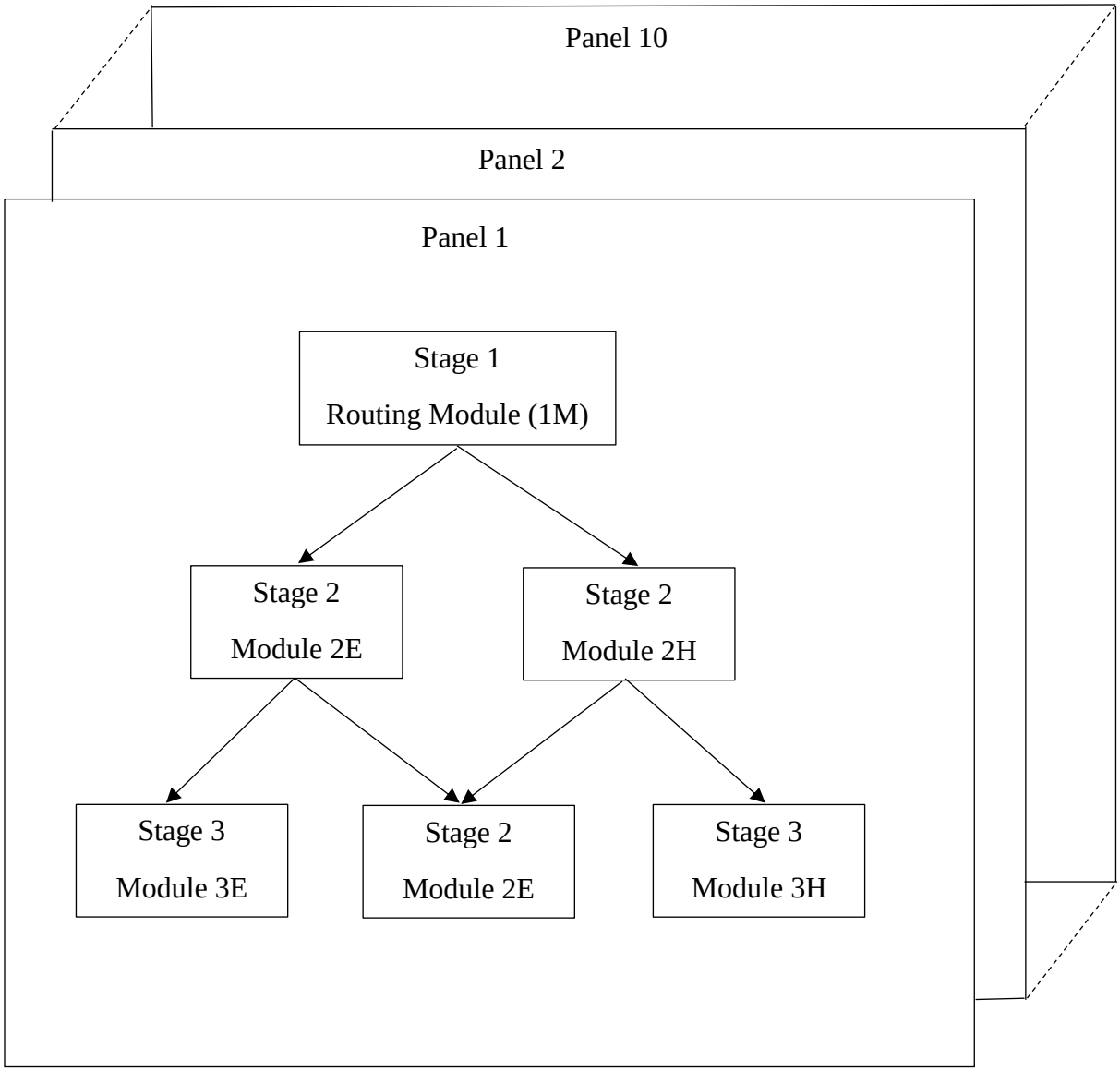


Figure 3.1 Example of 1-2-3 Multistage Test with 10 Panels. E = Relatively Easy; M = Moderately Difficult; H = Relatively Hard.

### 3.1.2 Test length

Test length is essential for measurement and scoring precision. Longer tests give more precise estimates but cost more in item writing and need longer administration time. Two levels of total test length were simulated in this study: moderate and long. Tests under the moderate length condition have 45 items. This is similar to the statewide assessment's length. Tests under the long test length condition have 60 items. This is equal to the averaged test length of four CPA

examination sections, 60 items. The number of items in a module was fixed as equal in each panel. Then, the corresponding number of items in each module in a panel is 15 and 20, respectively.

3.1.3. Item Pool Characteristics

The item pool should be designed carefully because the success of test assembly depends on the available items in the pool. As pointed by Van der Linden, Ariel & Veldkamp (2006), the item pool can be assembled from a set of fixed tests optimal at the distribution of examinee's ability level. Following this idea, the operational MST pool in this study was constructed based on each module considered as fixed tests. The three-parameter logistic model (3PLM) was used in generating the item pool designed for MST for generalization purposes. The probability of correct response in a 3PL model is defined as

$$P(X_i = 1|\theta) = c_i + (1 - c_i)\frac{\exp(Da_i(\theta - b_i))}{1+\exp(Da_i(\theta - b_i))} \quad (3.1)$$

where $a_j$ is the item discrimination parameter, $b_j$ is the item difficulty parameter, and $c_j$ is the pseudo-guessing parameter. Assuming this is a mathematics test covering five contents (e.g. Number, Data handling, Measurement, Algebra and Geometry), each content has an equal item number in the item pool.

Next, the current MST operational pool was cloned to obtain the master pool which should be guaranteed to have a sufficient large number of items meeting the content requirements. For convenience purposes, the master pool covered these five contents with equal number of items. Each item in the master pool was given a content code.

3.1.4 Data Generation

The data generation process included generating examinee's true ability, the operational item pool and the master pool. A group of 5000 examinees' true ability were randomly drawn from

the standard normal distribution. As indicated in Zheng et al.'s (2012) paper, the examinees' ability distribution was truncated within (-3.5, 3.5) to eliminate the effect of outliers. The values of the discrimination and guessing parameters in the MST pool were generated from the distributions, $a$ ~ lognormal (1, .3); $c$ ~ Uniform (.2, .1), to mimic a computerized test pool in mathematics (Leung, Chang & Hau, 2005). Discriminative power is a measure of item pool quality. The more discriminating items are selected, the more precise ability estimates will be. These distributions were selected to provide a high discriminating item pool to control item quality.

The distribution of item difficulty is essential on deciding target TIFs. The distribution of item difficulty parameters in the pool was generated to follow normal distribution to fit examinees' ability distribution. The desirable item difficulty level of each module aims to match the ability level of examinees routed to that module.

In the 1-2-3 design, item difficulty from stage 1 was designed to classify examinees accurately at the routing point into the next two modules. Since the examinee's ability distribution followed the standard normal distribution, roughly equal number of examinees tended to be routed to the easy and hard module in stage 2. Spray & Reckase (1994) suggested that choosing the most informative items at the cut-point yields efficient decisions. Thus, the mean item difficulty level of the module in stage 1 which is called the routing module, is supposed to be close to 0. Approximately, the item difficulty of the routing module followed the normal distribution N (0, 0.3) to ensure the item difficulty does not spread out from zero too much. Because of roughly equal number in module 2E and 2H, the item difficulty level was centered on -0.7 and 0.7 respectively. This was determined by the median ability level of two halves of examinees. In the last stage, three groups of examinees with nearly equal numbers were

desired to be routed to three modules 3E, 3M and 3H. Then, the mean difficulty of each module was set as -1, 0 and 1. The variance of item difficulty parameter should drop in later stages, as for narrowing down the region on the ability scale (Wang, 2013). To take account into the fact that the examinees are possibly routed to the wrong next module, the item difficulty distribution should have an overlap above a small range in the ability scale. The distributions of item difficulty parameter in each module are displayed in Table 3.2.

For the 1-3-3 design, the item difficulty level of the routing module followed the standard normal distribution N (0, 1) to match the ability level of examinees. This is different with the one of the 1-2-3 design because examinees were routed to three different modules rather than simple two modules. The similar process of determining item difficulty level was implemented in the stage 2 and 3 for the 1-3-3 design.

Table 3.2 Distributions for Item Difficulty Parameters in Each Module

| Design | 1-2-3 | 1-3-3 |
|---|---|---|
| Stage 1 | N (0,0.3) | N (0,1) |
| Stage 2E | N (-0.7, 0.6) | N (-1, 0.6) |
| Stage 2M | ---- | N (0, 0.6) |
| Stage 2H | N (0.7, 0.6) | N (1, 0.6) |
| Stage 3E | N (-1, 0.3) | N (-1, 0.3) |
| Stage 3M | N (0, 0.3) | N (0, 0.3) |
| Stage 3H | N (1, 0.3) | N (1, 0.3) |

Wang et al. (2012) noted that the MST pool size was set as 1.5 times of the number of items needed in most literature. In this study, there were a total of 9 modules with different item difficulty distributions. Based on the longer module length of 20 items and 10 panels in the 1-3-3 design, the ideal operational MST pool contained 2700 items for the two MST designs. Considering the real-life master pool size for AICPA, NCLEX and LSAT, the master pool size in this study was cloned from this operational MST pool 3 times to obtain 8100 items. Although the master pool was created by the MST pool, the large master pool size was able to guarantee the

fairness of drawing a CAT pool with a much smaller size than the MST pool. The information curve of the master pool is displayed in Figure 3.2. It follows the normal distribution which is similar to the information curve of LSAT's master pool. The information curve of the MST item pool can be found in the Appendix.

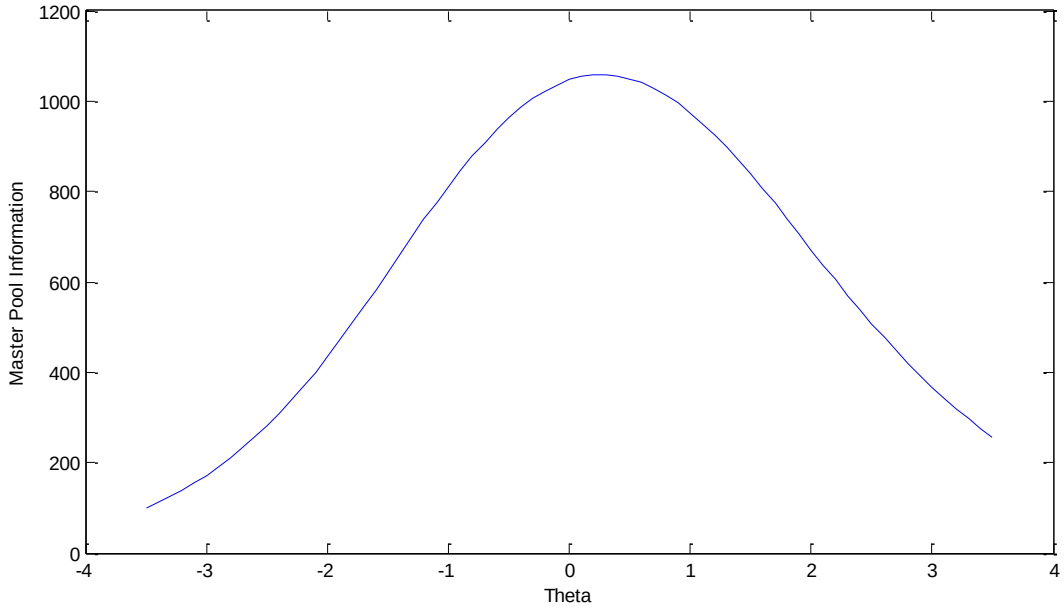


Figure 3.2 Information Curve of Master Pool

### 3.1.5 Test Assembly

#### 3.1.5.1 Test Information Function (TIF) targets

In this study, the Bottom-Up assembly strategy was employed to build the test. This strategy can assemble one or more versions of a panel simultaneously. Luecht et al. (2006) indicated that separate information targets and content constraints were required for each module under this strategy. For this reason, the TIF targets were specified in the module level by Approximate Maximum Information (AMI; Luecht, 2000) method. The AMI method was implemented by the following steps:

1. Specified a certain point on the ability scale with respect to the desired peak of TIFs for each module. For example, a predetermined point for the routing module is $\theta = 0$ in this study.

2. Computed the item information at that certain point (e.g., $\theta = 0$) in item pool.

3. Sorted the item information in the descending order.

4. Given the module length *n* and the number of panels *m*, the most informative $n \times m$ items were determined.

5. Computed the sum of item information of these $n \times m$ items at each of the selected points on the ability scale, $\theta_t$, $t = 1, \ldots, T$. (e.g., -3 to 3) with the increment of 0.1, then divided this amount by *m* to obtain the TIF targets. The target TIF is denoted as

$$\mathrm{TIF}(\theta_t) = \frac{\sum_{i=1}^{n*m} I_i(\theta_t)}{m} \tag{3.2}$$

Ten parallel modules without overlapped items across panels were built up simultaneously following the above procedures. Figure 3.3 displays an example of the target TIFs for each module in a forward-assembled 1-2-3 panel design. The examples of the target TIFs for each module in a backward-assembled 1-2-3 panel design, forward-assembled 1-3-3 panel design and backward-assembled 1-3-3 panel design can be found in the Appendix.

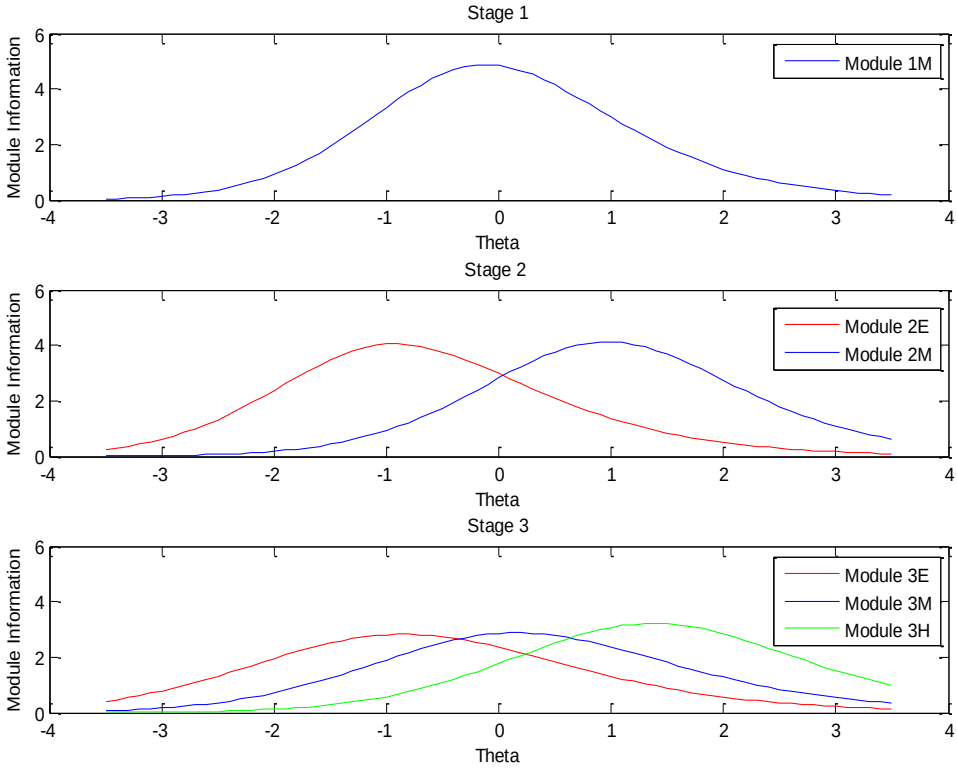


Figure 3.3 Module Level Target TIFs for One of the 1-2-3 Panels, Forward Assembly, 45 items

When maximizing the information at the same point for different modules (e.g. in "1-2-3" test design, routing module and medium-difficulty module have the target TIFs peaked at $\theta_t = 0$), two levels of assembly priority were considered here. The first one is forward assembly. As suggested by the term, modules in the early stage are built up prior to the ones in later stages. The other one is backward assembly, which means modules in later stage are built up prior to the ones in early stage. Table 3.3 lists the points where module information was maximized.

Table 3.3 The Points Where Module Information Was Maximized

| Design | "1-2-3" | "1-3-3" |
|---|---|---|
| Stage 1 | $\theta_t$ = 0 | $\theta_t$ = 0 |
| Stage 2 | $\theta_t$ = (-0.7, 0.7) | $\theta_t$ = (-1, 0, 1) |
| Stage 3 | $\theta_t$ = (-1, 0, 1) | $\theta_t$ = (-1, 0, 1) |

3.1.5.2 Assembly Algorithm

Luecht (2000) pointed out that the bottom-up strategy cannot be used if content constraints were not met in each stage. In addition to TIF targets as statistical constraints, separate content specifications as non-statistical constraints were required for each stage. Then, there was one content specification for each stage, respectively. For example, in 1-3-3 design, the content specification for three modules 2E, 2M and 2H are the same, even though they have different TIFs. Similarly, the three modules (3E, 3M and 3H) in Stage 3 meet the same content specification.

Once the target TIFs were determined, the normalized weighted absolute deviation heuristic (NWADH; Luecht, 1998) was used to build multiple panels simultaneously from the operational MST pool. This heuristic can handle any number and type of content or other categorical constraints. In the module having *n* items, as for identifying the *j*th item, this heuristic firstly computed the current information value by abstracting the value of selected items from the target value T($\theta_q$) ; then divided this value by the remaining number of items (*n*-*j*+1) . The expression in Equation 3.3 provides the target value of the next item to be selected.

$$\frac{T(\theta_q)-\sum_{i=1}^{j-1} I_j(\theta_q))}{n-j+1} \tag{3.3}$$

Secondly, the next item was selected with the information value matched to the value of Equation 3.3 in terms of all $\theta_q$ values. Then, the next item should maximize

$$e_i = 1 - \frac{\sum_{q=1}^{Q} d_{iq}}{\sum_{i\in R_{j-1}} \sum_{q=1}^{Q} d_{iq}} + \frac{c_i}{\sum_{i\in R_{j-1}} c_i}, \;\; i \in R_{j-1} \tag{3.4}$$

where $R_{j-1}$ represents index of the remaining items in the item pool except the selected *j*-1 items,

$$d_{iq} = \left| \left[ \frac{T(\theta_q)-\sum_{i=1}^{j-1} I_j(\theta_q) x_j}{n-j+1} \right] - I_j(\theta_q) \right| \tag{3.5}$$

where $x_j$ is a binary variable indicating whether each item was selected, and $c_i$ is the accumulated content weights for each unselected item in $R_{j-1}$

$$c_i = v_{ig}W_g + \left(1 + v_{ig}\right)\underline{W_g} \tag{3.6}$$

$$\underline{W_g} = W^{[max]} - \frac{1}{G}\sum_{i=1}^{G} W_g \tag{3.7}$$

The weights can be user-assigned integer weights. Adjusting Luecht's (1998) heuristics, the weights are given as followed:

$$\text{If} \quad \sum_i^{j-1} v_{ig} = Z_g, \quad then\ W_g = 0; \tag{3.8}$$

$$\text{If} \quad \sum_i^{j-1} v_{ig} < Z_g, \quad then\ W_g = 1; \tag{3.9}$$

Where $v_{ig}$ is the binary incidence of the item indicating a specific content constraint: $v_{ig}$ equals to 1 if the item belongs to the content constraint $g$, and equals to 0 if the item does not belong. $Z_g$ is the number of items required for each content in a module. For convenience, modules across stages were set to have the same content specifications in this study. For this reason, $Z_g$ was set as 3 for the module length of 15, and 4 for the module length of 20. User assigned proportional weights can be incorporated into the composition function in Equation 3.4 as well, in order to show the importance of satisfying statistical and non-statistical constraints (Luecht, 1998). Since this is a simulation study without real content specifications, user assigned proportional weights were not adopted.

Following the procedure of the Bottom-Up Strategy and NWADH, modules and panels were assembled successfully. Furthermore, to avoid the situation that panels assembled later tend to have worse-fitted items, items in each module and modules in each panel were permutated across panels to keep the quality of each panel equal. Examples of averaged module level information curves across panels in forward assembly and backward assembly for the 1-2-3 panel design are

displayed in Figure 3.4 and Figure 3.5, respectively. Each graph shows that the stages assembled earlier tend to have higher module level information. The figures of the 1-3-3 panel design and the figures of the 1-2-3 panel design with the test length of 60 items are presented in the Appendix. Figure 3.6 shows an example of the test information function for the routing modules (1M) across the ten backward assembled panels for the 1-2-3 panel design with test length of 45 items. Only a small variation ranged from the least to the most informative modules due to permutation. These three graphs were used to examine if the MST assembly process achieved its goals. Compared with the above Figure 3.3, the information curves in Figure 3.4 were a little lower due to the availability of items satisfying both information and content constrains.

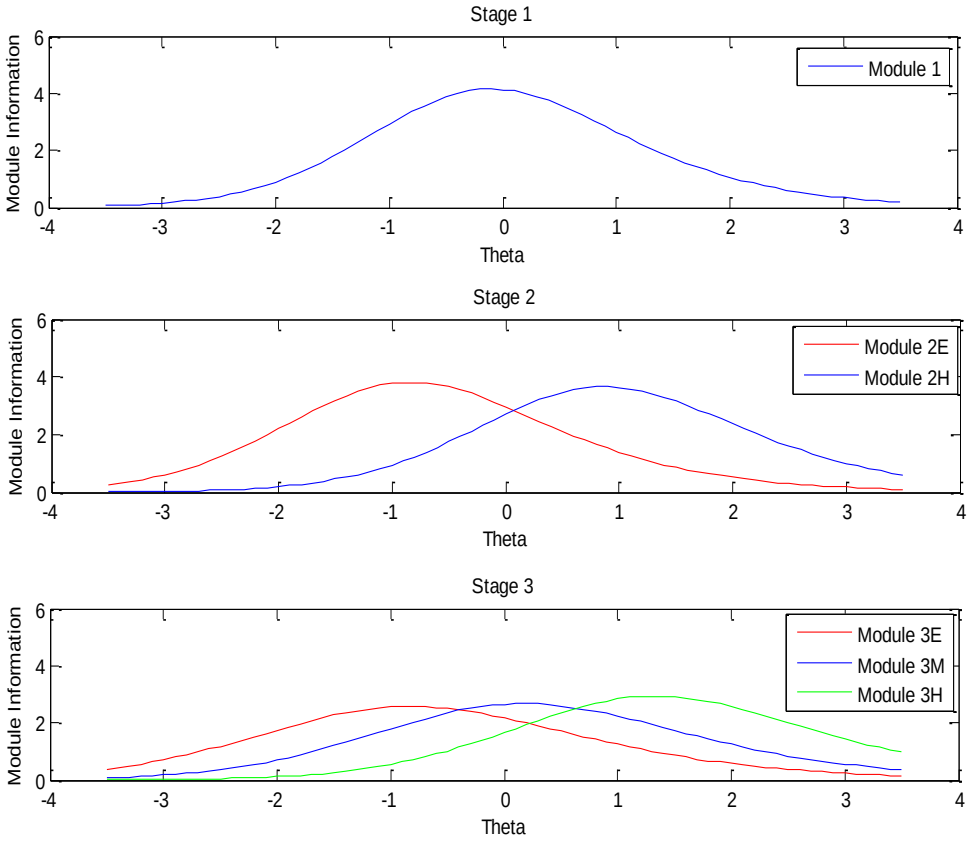


Figure 3.4 Averaged Module Level Information Curves across Forward Assembled Panels for the 1-2-3 Panel Design, 45 items

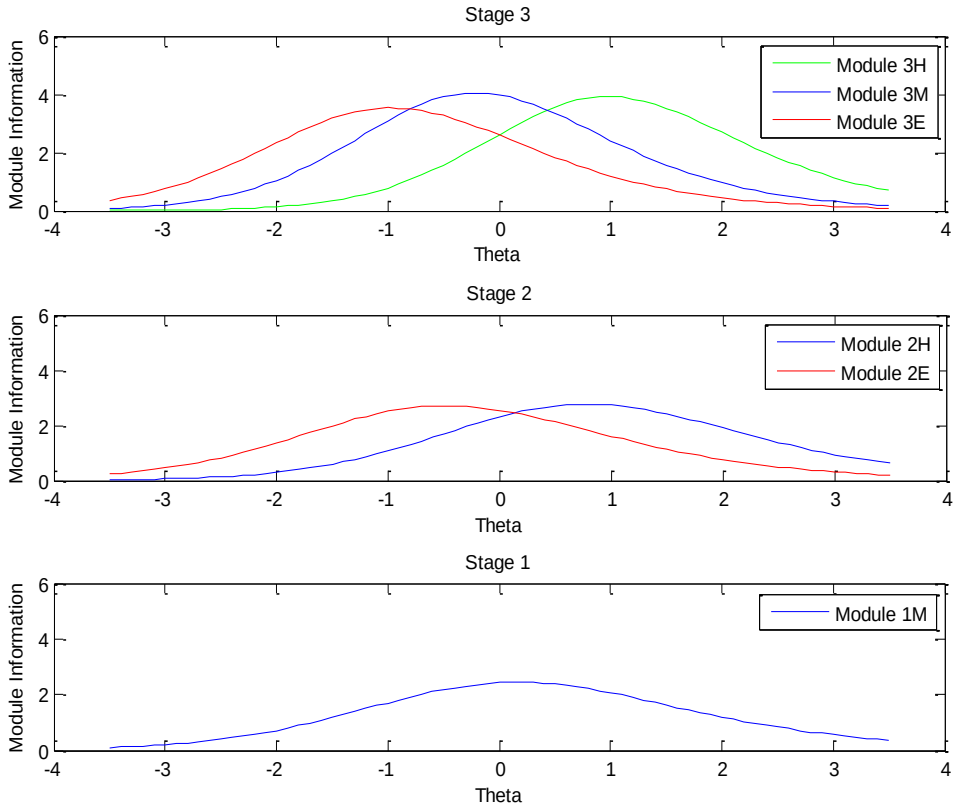


Figure 3.5 Averaged Module Level Information Curves across Backward Assembled Panels for the 1-2-3 Panel Design, 45 items

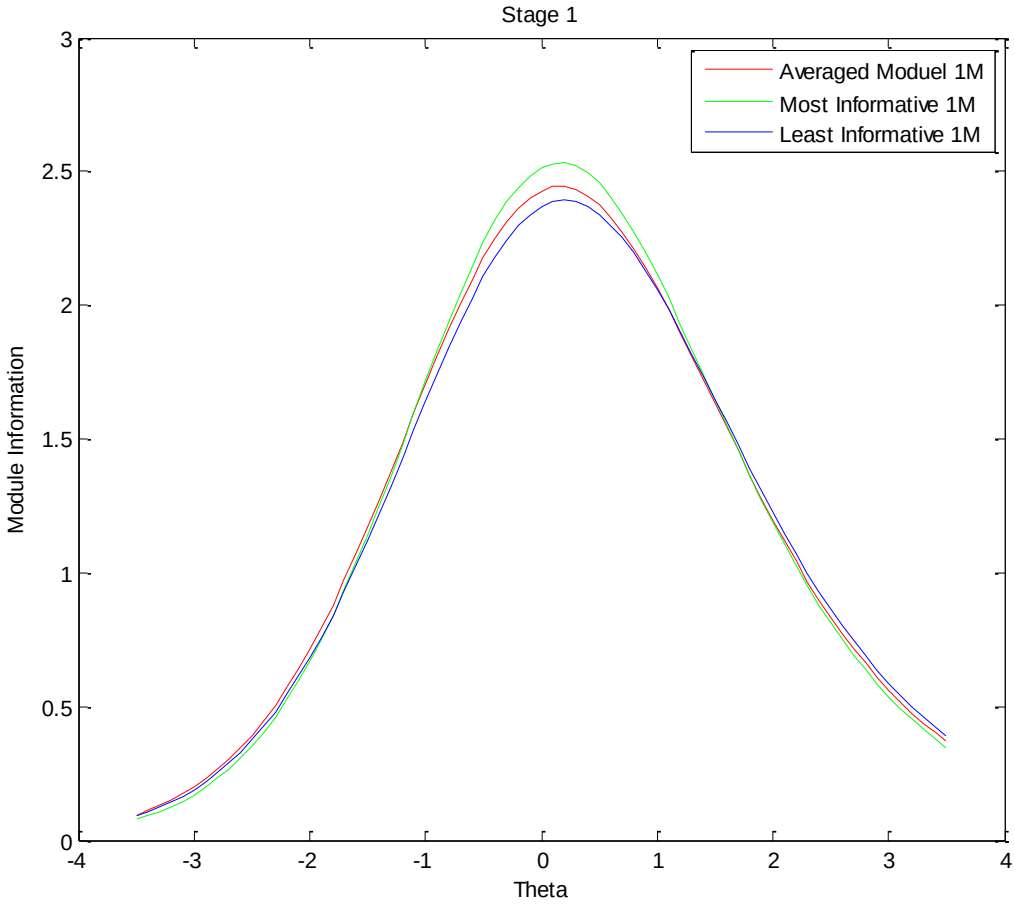


Figure 3.6 Averaged Module Level Information Curves of Module 1M across Backward Assembled Panels for the 1-2-3 Panel Design, 45 items (Green Graph = Most Informative Module 1M across 10 Panels; Blue Graph = Least Informative Module 1M across 10 Panels)

3.1.5.3 Routing Rules and Scoring

This study employed the approximate maximum information (AMI) method and the Defined Population Interval (DPI) to determine the routing points. Under the AMI method, the routing points were determined as the intersection point of the test information curves of the previous administered and current module (Luecht, Brumfield & Breithaupt, 2006). Figure 3.7 illustrates a routing procedure for AMI in a 1-3-3 panel design. Two routing points assigning examinees to stage 2 are determined in the 1-3-3 design, denoted as $\theta_L$ and $\theta_U$. The routing point $\theta_L$ corresponds to the intersection of test information curve of 1M+2E and 1M+2M, while the

routing point $\theta_U$ corresponds to the intersection of test information curve of 1M+2M and 1M+2H. Examinees with the estimated ability level below $\theta_L$ are assigned to the module 2E. Examinees with the estimated ability level above $\theta_U$ are assigned to the module 2H. Others are assigned to the module 2M. The routings points from stage 2 to 3, denoted as $\theta_L$, and $\theta_{U'}$, are determined by the summed test information curves over three stages. Since it is very likely to have different TIF of each module across panels, the routing points are likely to be different across panels as well.

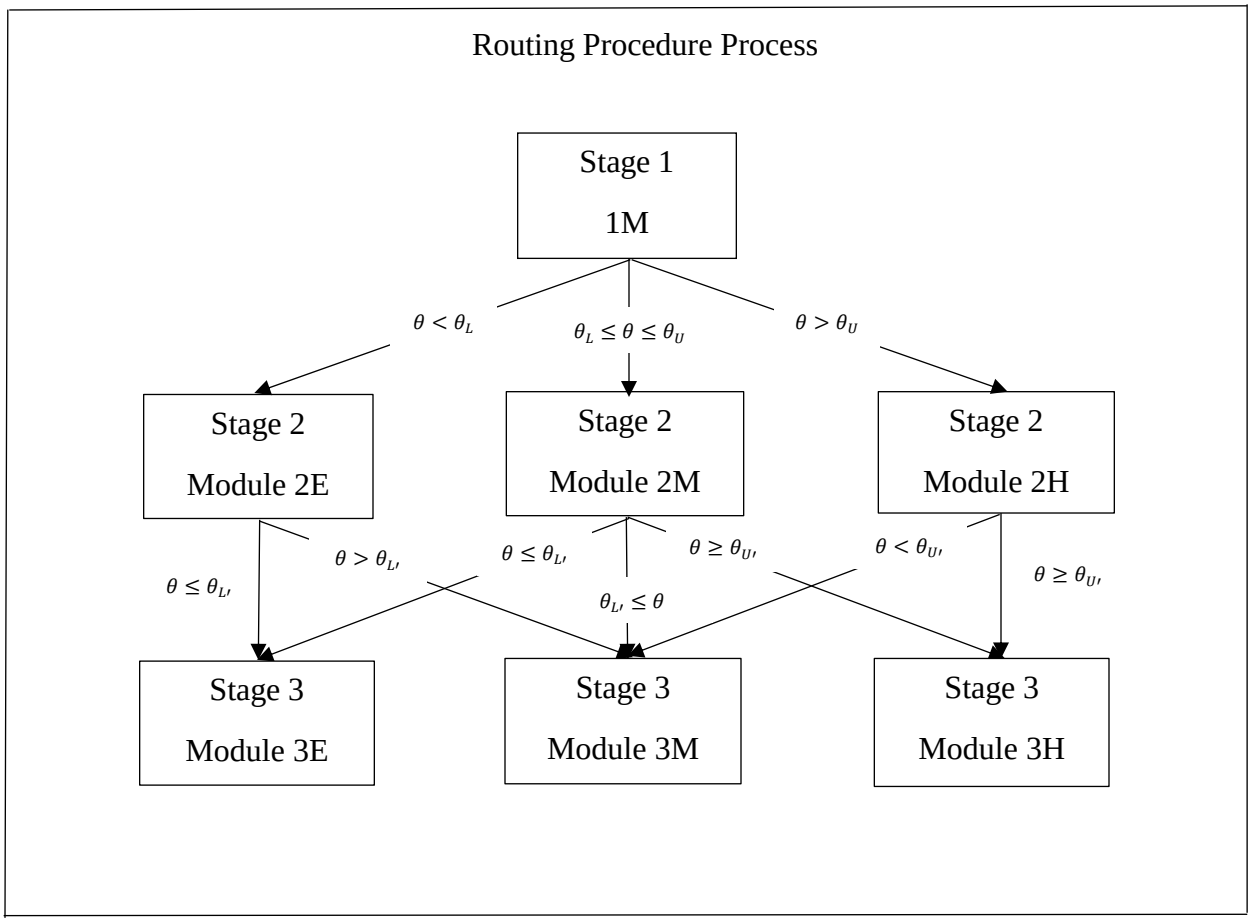


Figure 3.7 Example of AMI Procedure. AMI = Approximate Maximum Information. E = Relatively Easy; M = Moderately Difficult; H = Relatively Hard.

Under the DPI method, as a matter of policy, a pre-determined proportion of examinees is determined for each module in the next stage. In this study, roughly equal numbers of examinees were expected in each module of stage 2 and 3. That is to say, the 50th percentile is the routing point from stage 1 to stage 2 in the 1-2-3 design, while the 33rd and 67th percentiles are the routing points in the 1-3-3 design and in the transition from stage 2 to 3 of the 1-2-3 design. Since the distribution of examinees followed normal distribution, the routing points were -0.43 and 0.43 for the 1-3-3 design, and 0, -0.43 and 0.43 for the 1-2-3 design. After identifying the routing points, the scoring procedure was implemented. As mentioned earlier, this study applied 3PL-IRT model for all dichotomously-scored items in the whole simulation procedure.

3.1.6 Test Administration

For example, the 1-3-3 MSTs were administrated in this study as the following steps after the modules and panels.

1. The examinee was randomly assigned one of the ten panels.
2. The examinee would take the routing module (i.e., Module 1M with medium difficulty level) of the assigned panel.
3. And the end of the routing module, the ability estimate ($\hat{\theta}$) of the examinee was obtained by MLE.
4. The examinee was routed to a module (2E, 2M or 2H) of stage 2 based on the comparison of $\hat{\theta}$ and predetermined routing points from stage 1 to 2.
5. And the end of the routing module, the updated ability estimate ($\widehat{\theta'}$) of the examinee was obtained by MLE.
6. The examinee was routed to a module (3E, 3M or 3H) of stage 3 based on the comparison of $\widehat{\theta'}$ and predetermined routing points from stage 2 to 3.

7. After stage 3, the final ability estimate of the examinee was obtained and recorded. The test information was calculated and recorded as well for the comparison with CAT.

## 3.2 CAT Simulations

### 3.2.1 Item Pool Characteristics

An important characteristic of the item pool is to have enough items to cover the whole ability scale. The CAT item pool should have enough items as well to give a fair comparison. Guidelines for the appropriate size of the item pool are from six to twelve times the test length (Weiss, 1985; Stocking 1994; Gu, 2007). In order to match the maximum item exposure rate of 0.1 in MSTs which is strict in CATs, this study adopted 15 as the ratio of pool size to test length. Since the long test has 60 items, the pool size was set as 900 which also guarantees enough items for the moderate test length of 45. There were equal numbers of items in 5 contents. In this study, the operational CAT pool consisting of 900 items will be constructed from the master pool by following the *a*-stratified method (Chang & Ying, 1999) and "item distribution for the .96-optimal item pool with exposure control" (Mao, 2014, p. 63). The later one illustrated the number of items in *b*-bin in a .96-optimal CAT item pool. The definition of ".96-optimal item pool" represents an item pool "that always has an item available for selection that *p*% matches the desired characteristics specified by the item selection routine for the CAT" (Reckase, 2007). Given the item distribution for such an optimal item pool, the procedure of CAT item pool construction had the following steps:

1. Divided the master pool into 5 sub-pool by content, then each content had equal number of items (i.e., 1620) as well.
2. Sorted the items in each sub-pool based on an ascending order of the *a*-parameters.

3. Partitioned each sub-pool into 3 strata by *a*-parameter with lowest-*a* items being put in the first stratum and largest-*a* items being put in the last stratum. Each stratum has equal number of items (i.e., 540).
4. Within each *a*-stratum, the number of items in the *b*-bin (e.g. $b < -1.7$, $-1.7 \leq b < -0.2$; $-0.2 \leq b < 0.2$; $0.2 \leq b < 1.7$; and $b \geq 1.7$) were determined by following the item distribution for the .96-optimal item pool with exposure control. The current study adopted more examinees and lower item exposure rate than the ones of Mao's study (2014). For this reason, more items were needed for the examinee whose ability level was near 0 on the ability scale. Table 3.4 illustrates the distribution of item frequencies within each *b*-bin.
5. Randomly drew items following the item distribution shown in Table 3.4 within each stratum.
6. All items drawn in each stratum were pooled across sub-pools to finalize the CAT operational item pool. The example of item distribution for the CAT item pool is shown in Figure 3.8 below.

Table 3.4. Item Distribution for Each *a*-Stratum

| | $b < -1.7$ | $-1.7 \leq b < -0.2$ | $-0.2 \leq b < 0.2$ | $0.2 \leq b < 1.7$ | $b \geq 1.7$ |
|---|---|---|---|---|---|
| Number of items | 8 | 18 | 7 | 18 | 8 |

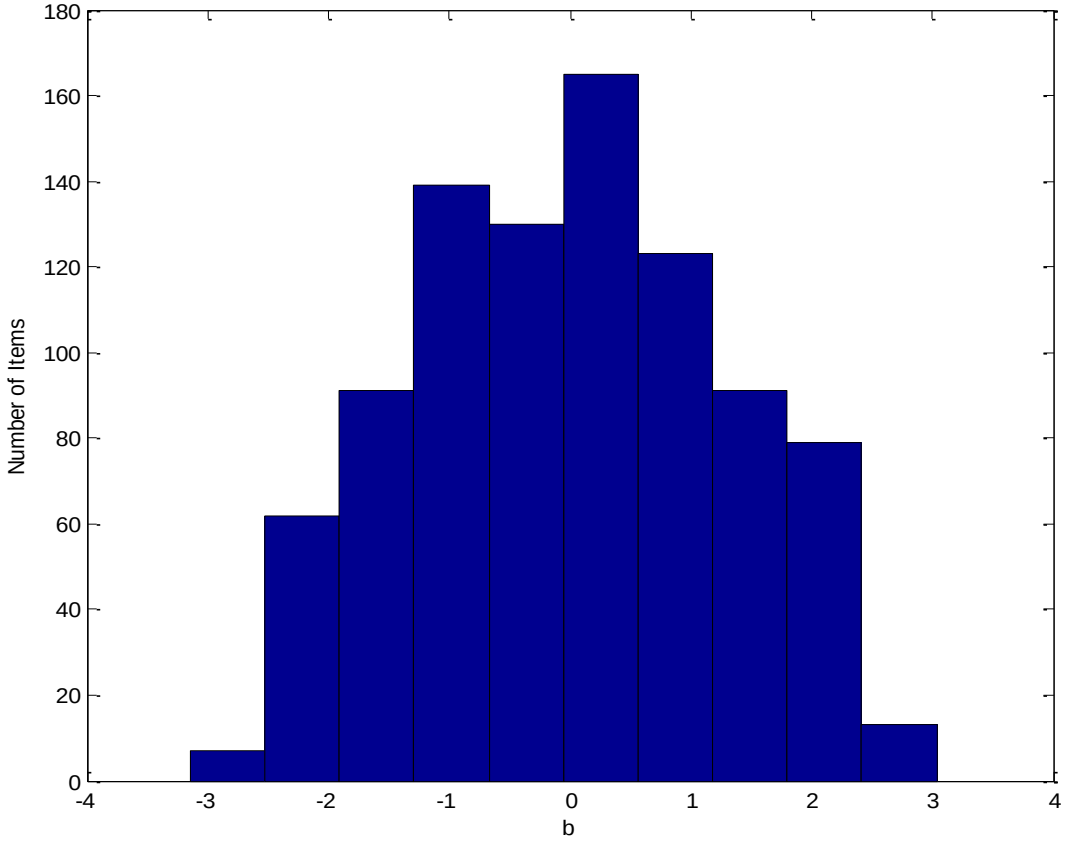


Figure 3.8 Example of Distribution of Difficulty (*b*) Parameter in the CAT Item Pool

### 3.2.2 Item Selection Procedure

When incorporating multiple statistical constraints and test requirements simultaneously in CATs, the item selection procedure becomes more complicated than the basic procedure. This study employed the Maximum Priority Index (MPI; Cheng & Chang, 2009) method which is very promising for constrained item selection in CAT. It successfully controls multiple constraints (e.g. exposure control and content constrains) simultaneously, requires no weight adjustment and can be applied with different item selection algorithm. The MPI is computed for each eligible item after administering one item. The item with larger MPI tends to be selected more likely. The priority index is denoted:

$$PI_{i_c} = I_{i_c} \prod_{k=1}^{K} (w_{kc} f_k)^{c_{ik}} \quad (3.10)$$

where $I_{i_c}$ represents Fisher information of item $i_c$ at current ability estimate; $i_c$ is the number of items in the pool; $c_{i_c k}$ equals 1 when item l is relevant to constraint *k*, otherwise it equals to 0; $f_k$ represents the scaled "quota left" of constraint *k* and $w_{kc}$ represents the weight according to $f_k$ in CAT simulation. For convenience purpose, the weight of the constraint *k* ($w_{kc}$) was fixed as 1 in this study. When constraint *k* represents content constraint, the scaled "quota left" is computed as

$$f_k = \frac{(X_k - x_k)}{X_k} \tag{3.11}$$

where $X_k$ represents the number of items required from a certain content area, and $x_k$ represents the number of items have been selected. Due to the property of variable length CAT, there is no specific number of items required from each content for all examinees. This study assigned a value to $X_k$ by maximum number of items from each content. Followed by equal number of items across 5 contents, $X_k$ was fixed as 11 for the maximum test length of 55 items, and as 14 for the maximum test length of 70 items. When constraint *k*' represents exposure control constraint and require the item exposure rate to be no more than *r*, the scaled "quota left" is computed as

$$f_{k'} = \frac{(r - (\frac{n_e}{N_e}))}{r} \tag{3.12}$$

where $N_e$ represents the total number of examinees, and $n_e$ represents the number of examinees who have seen item l. Because the items in MSTs had item exposure rate no more than 0.1, *r* in the above equation (3.12) was fixed as 0.1.

3.2.3 Data Generation

Ten replications were implemented in CAT simulation as well. In the CAT simulation, the initial value of examinees was randomly generated from the uniform (-0.4, 0.4) distribution. For fair comparison purpose, there were four factors matched with those of MST simulations,

including similar conditional test information, IRT scoring procedure, maximum item exposure rate and content specifications. In this variable-length CAT, the test was terminated once the test information conditional on an examinee's ability level fell into the range from below to above 5% of the corresponding conditional test information of MST. Even though the goal is to make a fair comparison, it was not realistic to have exactly the same conditional standard error of measurement. The test was terminated when the test length reached 55 items for the test length of 45 in MST, and 60 items for the test length of 70 in MST. This was to avoid an endless test for the examinees having extreme ability level. The IRT-based score produced by Maximum Likelihood Estimation (MLE) was used in item selection procedure of CAT administration, and a final IRT-based score was reported to the examinees.

### 3.3 Evaluation Criteria

The performances of MST and CAT were evaluated based on ability estimates. The evaluation criteria for ability estimate including mean of bias, and mean squared error (MSE). The mean bias was calculated as

$$Mean\ Bias = \sum_{e=1}^{N_e} \frac{\hat{\theta}_e - \theta_e}{N_e}, \tag{3.13}$$

where $N_e$ was the number of examinees, $\hat{\theta}_e$ was the estimated ability, $\theta_e$ was the true ability. And MSE was obtained by Equation 3.14

$$MSE = \sum_{e=1}^{N_e} \frac{(\hat{\theta}_e - \theta_e)^2}{N_e}, \tag{3.14}$$

Considering the testing mode change from CAT to MST, the comparison of MST and CAT was evaluated by averaged test length for item writing cost and test administration time, as well as measurement accuracy.

The first analysis dealt with the question of which MST design gave the highest measurement precision under different conditions. Further, whether MST outperformed CAT in terms of

measurement accuracy was examined. The mean bias and MSE were calculated for both the corresponding CAT and MST designs under each of 16 conditions. Each condition of MST was compared with the corresponding result of CAT.

The other research question focused on the comparison between averaged test length of CAT and fixed-length MST. When considering testing mode change, administration time, measurement accuracy and item cost should be noticed. Which testing mode gave a shorter test based on a fair comparison was examined.

# CHAPTER 4 RESULTS

The results for CAT and MST designs are presented in this chapter. They were compared on measurement accuracy, and these findings are presented first. They are followed by the comparison of averaged test length in MST and the corresponding CAT for the purpose of testing mode change. This study also found a situation of routing point shift using AMI routing strategy for backward assembled MSTs, which is presented next. The Maximum Priority Index (MPI) was used to select items in the variable-length CAT of this study. The results of content balancing are shown last. MST designs were simulated across four factors: panel design, test length, routing strategy and assembly priority. All tables and figures are included in this chapter. All results are averaged across the ten replications under each condition.

## 4.1 Measurement Accuracy

Measurement accuracy was evaluated by the degree of ability estimate recovery. The overall results included mean bias and mean squared error (MSE). These were compared across MST conditions first, and between each condition of MST and the corresponding CAT.

Table 4.1 to 4.4 give the overall measurement accuracy statistics for CAT and MST across all conditions. These tables show that the overall measurement accuracy was good. Table 4.1 indicates that forward assembled MSTs always have slightly higher mean biases than backward assembled MSTs for the moderate length test (i.e., 45 items). The absolute difference ranged from .002 to .011. No obvious trend in panel designs and routing strategies was found among forward assembly conditions. Under backward assembly condition, similar mean biases were obtained between MST panel designs and routing strategies. All of them are close to 0, even though backward assembled MST using DPI routing strategy always underestimated examinees' abilities slightly. Compared

to CAT, MST always provided slightly smaller mean biases, which absolute difference ranged from .006 to .025.

Table 4.1 Mean Bias of the Estimated $\theta$ for Moderate Length Test

| Test length | Testing Mode | Design | Forward AMI | Forward DPI | Backward AMI | Backward DPI |
|---|---|---|---|---|---|---|
| 45 | CAT | Item-Level | .018 | .022 | .025 | .023 |
| | MST | 1-2-3 | .012 | .003 | .001 | -.001 |
| | CAT | Item-Level | .019 | .020 | .020 | .024 |
| | MST | 1-3-3 | .004 | .011 | .001 | -.001 |

Note: All statistics were computed across 10 replications; each replication has 5,000 examinees.

Table 4.2 indicates the mean biases of the long test (i.e., 60 items) across all conditions. The results demonstrate that there is no obvious difference in mean bias between assembly priorities, panel designs and routing strategies. CATs have similar mean bias with MSTs. When comparing the mean bias between two test length conditions, it is shown that the longer test had a slightly smaller mean bias. Figure 4.1 displays a straightforward comparison across all MST conditions.

Table 4.2 Mean Bias of the Estimated $\theta$ for the Long Test

| Test length | Testing Mode | Design | Forward AMI | Forward DPI | Backward AMI | Backward DPI |
|---|---|---|---|---|---|---|
| 60 | CAT | Item-Level | .008 | .015 | .015 | .018 |
| | MST | 1-2-3 | .005 | .007 | .007 | .006 |
| | CAT | Item-Level | .013 | .013 | .004 | .011 |
| | MST | 1-3-3 | .007 | .011 | .004 | .002 |

Note: All statistics were computed across 10 replications; each replication has 5,000 examinees.

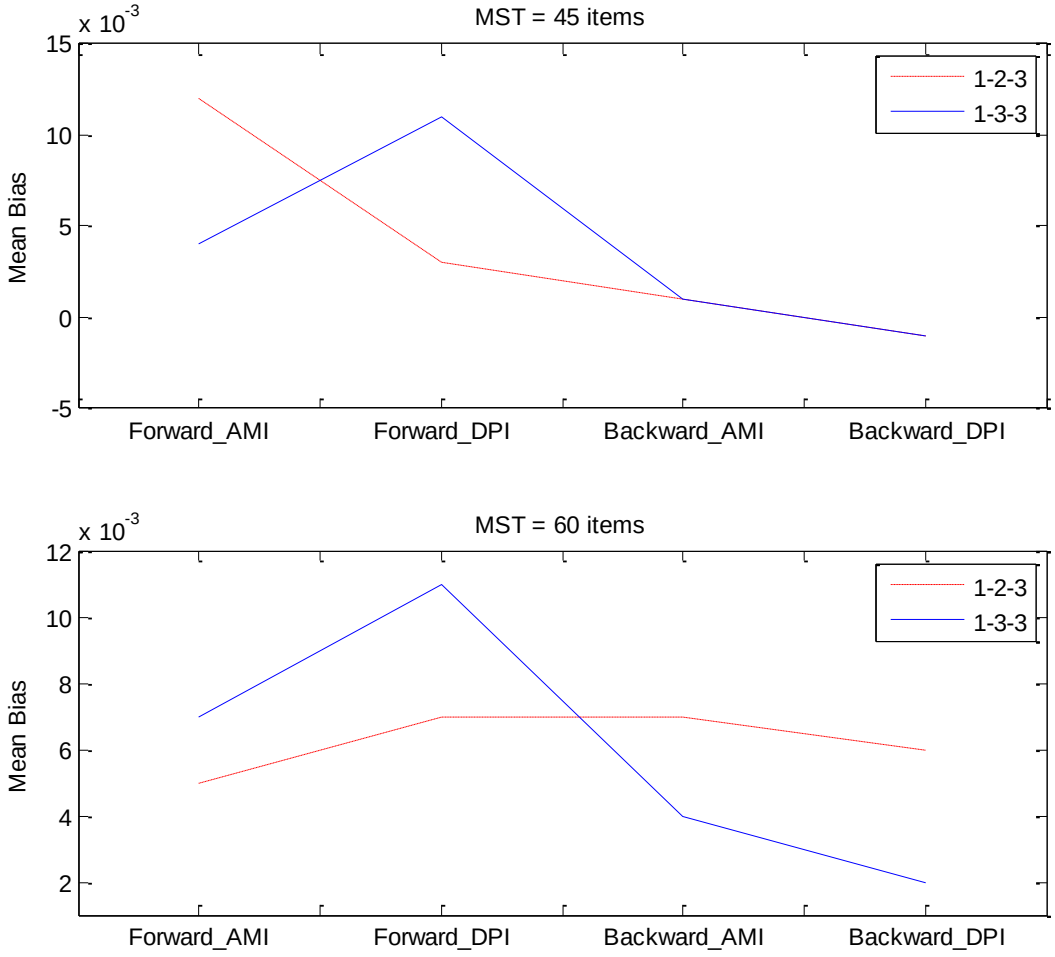


Figure 4.1 The Mean Biases under Different MST Design Conditions

Table 4.3 and 4.4 list the MSE of the estimated ability for the moderate-length test and long test, respectively. With respect to the moderate-length test exhibited by Table 4.3, the MSEs were smaller in MSTs than in CATs, whose differences ranged from .02 to .03. There is a slight difference of .01 and .02 in MSEs between forward and backward assembled MSTs. The routing strategies and panel designs provided similar MSE.

Table 4.3 MSE of the Estimated $\theta$ for Moderate Length Test

| Test length | Testing Mode | Design | Forward AMI | Forward DPI | Backward AMI | Backward DPI |
|---|---|---|---|---|---|---|
| 45 | CAT | Item-Level | .14 | .15 | .16 | .16 |
| | MST | 1-2-3 | .12 | .12 | .14 | .14 |
| | CAT | Item-Level | .15 | .15 | .15 | .16 |
| | MST | 1-3-3 | .13 | .13 | .13 | .14 |

Note: All statistics were computed across 10 replications; each replication has 5,000 examinees

The patterns of MSE were similar in both the moderate-length and long test. As indicated by Table 4.4, MSEs are slightly higher in backward than forward assembled MSTs. The routing strategies and panel designs did not have an impact on the MSE. It is noted than MSTs gave slightly smaller MSEs than CATs, ranging from .01 to .02. According to the comparison between test lengths, a longer test has a smaller MSE. Figure 4.2 displays a straightforward comparison among MST conditions.

Table 4.4 MSE of the Estimated $\theta$ for Long Test

| Test length | Testing Mode | Design | Forward AMI | Forward DPI | Backward AMI | Backward DPI |
|---|---|---|---|---|---|---|
| 60 | CAT | Item-Level | .12 | .12 | .13 | .13 |
| | MST | 1-2-3 | .10 | .10 | .11 | .11 |
| | CAT | Item-Level | .12 | .12 | .14 | .13 |
| | MST | 1-3-3 | .10 | .10 | .12 | .12 |

Note: All statistics were computed across 10 replications; each replication has 5,000 examinees.

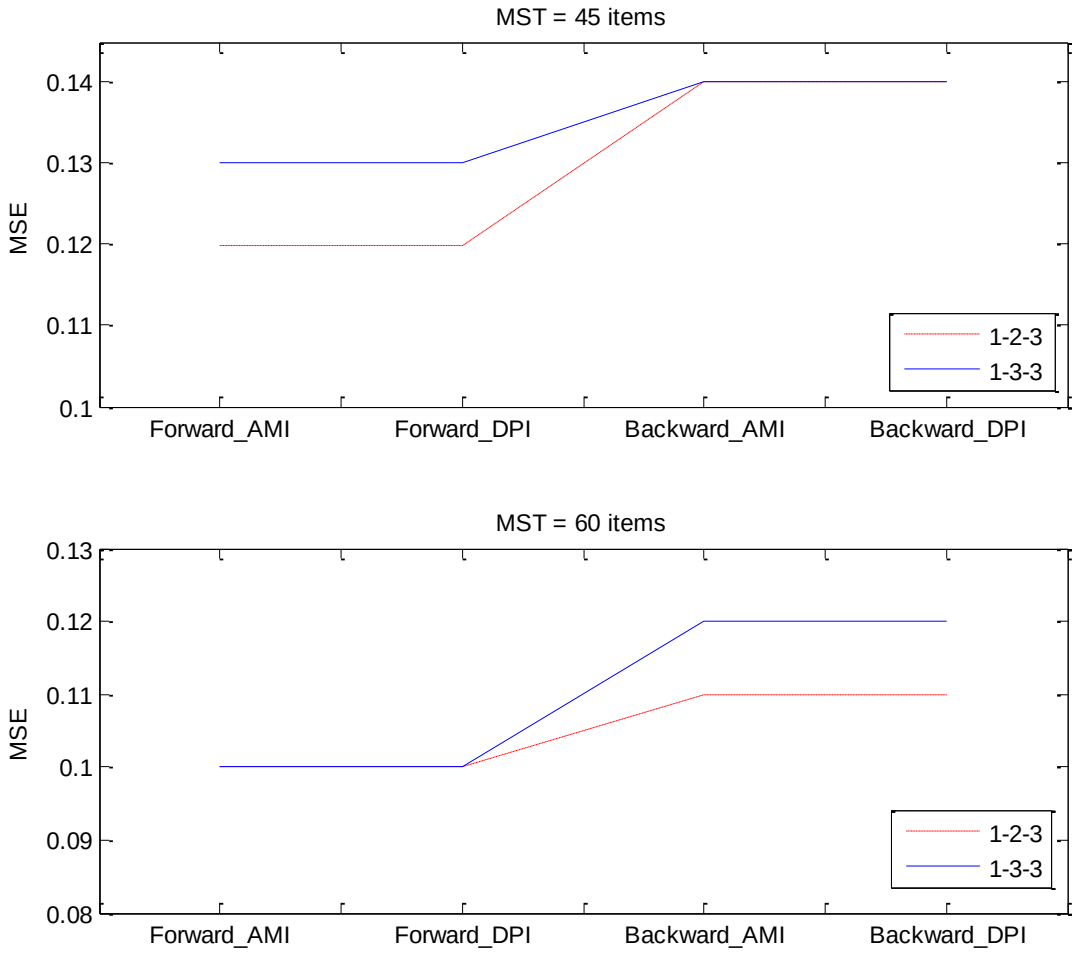


Figure 4.2 The MSEs under Different MST Design Conditions

In summary, Tables 4.1, 4.2, 4.3 and 4.4 demonstrated that the difference in overall measurement accuracy across MST conditions was very small. Different panel designs, assembly priorities and routing strategies did not have a considerable impact on measurement accuracy. Longer tests performed slightly better than short tests. Measurement accuracy of the MST designs was slightly higher than the corresponding CATs on the overall ability scale.

## 4.2 Averaged Test Length

Table 4.5 summarizes the average test length in CAT matched with MST conditions. No obvious difference in average test length was indicated by the comparison between routing strategies. Consistently, the averaged test lengths in CAT corresponding to backward assembled MSTs were shorter than those with forward assembled MSTs. The difference between forward and backward assembly increased as test length increased. For comparing the average test length between MST and the corresponding CAT, it is noted that there was a slight difference of 1 or 2 items under forward assembly conditions for the moderate-length test. In contrast, a large difference of 5 or 6 items was noticed under backward assembly conditions. For the long test, there was a difference of 3 to 6 items between the MST and the corresponding CAT under forward assembly conditions, and a difference of 9 to 11 items under backward assembly conditions. With respect to panel designs, both 1-2-3 and 1-3-3 panel designs required similar averaged test length in CAT for the moderate-length test. But for the long test, the 1-3-3 panel design needed shorter averaged test length in CAT than the 1-2-3 panel design did, whose difference ranged from 1 to 3 items. Figure 4.3 shows a clear comparison. A plausible explanation is presented in Chapter 5.

Table 4.5 Average Test Length in CAT with the Corresponding MST Conditions

| Test length | MST Panel Design | Forward AMI | Forward DPI | Backward AMI | Backward DPI |
|---|---|---|---|---|---|
| 45 | 1-2-3 | 44 | 44 | 40 | 40 |
| | 1-3-3 | 44 | 43 | 40 | 39 |
| 60 | 1-2-3 | 57 | 57 | 51 | 51 |
| | 1-3-3 | 55 | 54 | 50 | 49 |

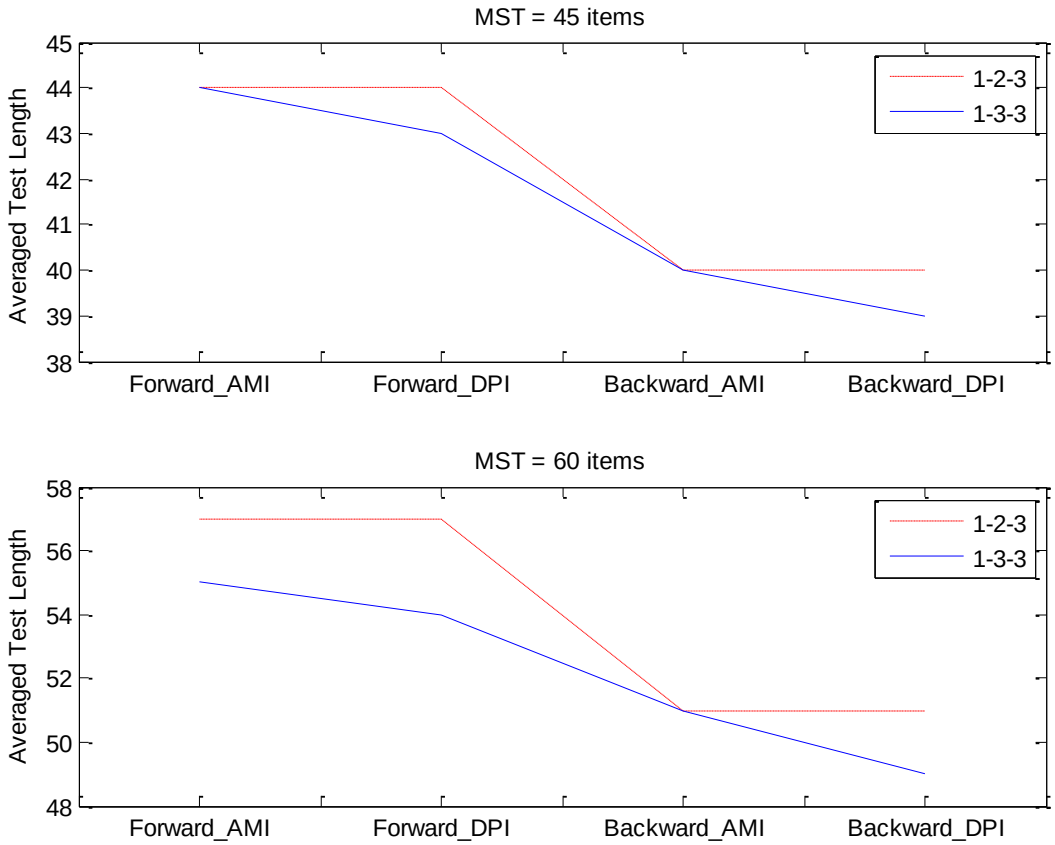


Figure 4.3. The Averaged Test Lengths of CAT with the Corresponding MST Conditions

### 4.3 Routing Point Shift in MSTs using AMI routing strategy

In addition to the measurement accuracy and average test length, this study also found a situation that routing points shifted in backward assembled MSTs using AMI strategy. Tables 4.6 and 4.7 show the average percentage of examinees routed to each module in the 1-2-3 and 1-3-3 panel design over 10 replications, respectively. For both the moderate-length and long test, the patterns of routing point shift were the same. In Table 4.6, for both forward and backward assembled MSTs, the amount of examinees routed to Module 2E and 2H were almost half and half. In stage 3, the amount of examinees routed to Module 3M was the largest in comparison to Module 3E and 3H in both forward and backward assembled MSTs. There are over 30 percent of

the examinees taking Module 3E, and over 20 percent of the examinees taking Module 3H in the forward assembled MSTs. In contrast, 20 percent of the examinees taking Module 3E and almost 40 percent examinees taking Module 3H in the backward assembled MSTs. This result indicated that the routing points in stage 3 shifted to the left in the backward assembly condition.

Table 4.6 The Averaged Percentage of Examinees Routed to Each Module in the 1-2-3 Panel Design over 10 Replications (%)

| Module | Test length = 45 | | Test length = 60 | |
|---|---|---|---|---|
| | Forward | Backward | Forward | Backward |
| 2E | 51 | 48 | 52 | 50 |
| 2H | 49 | 52 | 48 | 50 |
| 3E | 32 | 21 | 35 | 19 |
| 3M | 42 | 41 | 42 | 44 |
| 3H | 27 | 38 | 23 | 37 |

For the 1-3-3 panel design summarized by Table 4.7, the largest amount of examinees in stage 2 were routed to module 2E under forward assembly conditions. This is similar with the examinee distribution in stage 2 in the 1-2-3 panel design. Module 2M and 2H divided the remaining examinees, which only took nearly 5 percent examinees (i.e., 52 percent on Module 2E in the 1-2-3 panel design – 47 percent on Module 2E in the 1-3-3 design = 5 percent) from Module 2E. With respect to moderate-length versus long test under forward assembly condition, the amount of examinees taking Module 3E was the largest, followed by Module 3M, and then Module 3H.

Compared to forward assembly condition, the routing points from stage 1 to 2 under backward assembly condition shifted to the left to move 20 percent examinees from Module 2E to the remaining two modules. Only 27 and 26 percent examinees were routed to Module 2E in moderate-length and long test, respectively. The routing points from stage 2 to 3 shifted to the left even more compared to the ones from stage 1 to 2. Nearly 20 percent examinees were routed to Module 3E. Similar numbers of examinees took Module 3M and 3H.

Table 4.7 The Average Percentage of Examinees Routed to Each Module in the 1-3-3 Panel Design over 10 Replications (%)

| Module | Test length = 45 | | Test length = 60 | |
|---|---|---|---|---|
| | Forward | Backward | Forward | backward |
| 2E | 47 | 27 | 47 | 26 |
| 2M | 20 | 39 | 25 | 37 |
| 2H | 33 | 34 | 28 | 37 |
| 3E | 40 | 22 | 41 | 19 |
| 3M | 33 | 37 | 33 | 41 |
| 3H | 27 | 41 | 25 | 40 |

In summary, for the 1-2-3 panel design, the routing points from stage 2 to 3 shifted to the left in backward assembly condition compared with forward assembly condition. Over 10 percent fewer examinees were routed to Module 3E in backward rather than in forward assembled MST. The number of examinees routed to Module 3M was similar.

The situation for the 1-3-3 panel design was more complex than the 1-2-3 panel design. According to the forward assembly condition, the largest number of examinees (i.e., 47 percent) were routed to Module 2E, instead of the expected Module 2M. In contrast, the number of examinees on Module 2E was the smallest in stage 2 under backward assembly condition. Module 2M and 2H had similar number of examinees. Since the extreme pathway is unlikely to happen, Module 3E and 3M kept similar amounts as well. Compared to the forward assembly condition, the routing point from the routing module to 2E shifted to the left under the backward assembly condition. From stage 2 to 3, the routing points shifted further to the left. A plausible explanation is presented in Chapter 5 below.

### 4.4 Content balance using MPI in a variable-length CAT

Previous research employed MPI in the fixed-length CAT. This study attempted to employ it in a variable-length CAT and to examine the content balancing. As illustrated by Equation 3.11, the maximum number of items in each content was set to be one fifth of the maximum test

length, since an equal number of items was required in five contents. Under each condition over ten replications, more than 90 percent of examinees had the equal number of items or a difference of 1 item across contents. This slight content violation is acceptable.

# Chapter 5 DISCUSSION and CONCLUSION

This chapter provides a summary, a discussion and limitations of the results. It has four main sections. The first section summarizes the research objectives, the methodology applied in this study, and the results. The discussion of the results based on the major finding are then described, followed by implications. Limitations of this study and directions for future research are discussed in the final section.

## 5.1 Summary of This Study

The main purposes of this study were (1) to compare the measurement accuracy of MSTs with the corresponding CATs; (2) to investigate which MST designs will give the highest measurement accuracy under different conditions (e.g., 1-2-3 and 1-3-3 panel designs, forward and backward assembly, routing strategies of AMI and DPI and the test lengths of 45 and 60 items); and (3) to compare which testing mode (i.e., CAT or MST) will give a shorter test under each item pool matching similar test information on overall ability scale, item exposure rate and test content specification. The goal of this study was to make a fair comparison which includes three facets: (1) creating an item pool for MST and CAT respectively; (2) matching similar conditional test information for examinees; and (3) providing other matched properties during test administration (i.e., similar level of availability of items in each item pool, maximum item exposure rate, and content specifications). Then, a master pool was created to ensure the availability of items for maintaining MST item pool and drawing CAT item pool. A simulation study was conducted to build a MST item pool first, and then to assemble MST panels based on target TIFs in each module according to different panel designs, test lengths and assembly priority. The MSTs were administrated across different conditions. The comparison between

MST and CAT were on both measurement accuracy and then on test length when considering testing mode change. In addition to the research objectives, the shifted routing point in backward assembled MSTs using AMI routing strategy were found in this study, as well as the examination of the MPI in a variable-length CAT.

### 5.1.1 Measurement Accuracy Criteria

No meaningful difference in mean bias of ability estimate was found among different conditions of test length, of routing strategy, of assembly priority and of panel design. Generally, the mean biases were all close to 0.

In terms of MSE, no notable difference was demonstrated among different conditions of routing strategy, of assembly priority and of panel design. Among all conditions, the MSE decreased as test length increased. But the magnitude was small, ranging from .01 to .03.

With respect to the overall measurement accuracy across MST conditions, there was no difference between routing strategies, panel designs and assembly priorities. Long test performed slightly better than moderate-length test.

Comparing measurement accuracy between all conditions of MSTs and the corresponding CATs, MSTs outperformed CATs slightly and consistently in both MSE and mean bias. The magnitude of difference only ranged from 0 to .025 across all conditions in mean bias, and from .01 to .02 across all conditions in MSE.

### 5.1.2 Test Length Criteria

Based on the fair comparison between MST and CAT in this study, CAT provided a slightly shorter test than the corresponding forward assembled MSTs. When MST had a fixed test length of 45 items, CAT can achieve similar conditional test information with MST by a test length having one or two items less in average than MST. The small difference of 1 item was noted

between panel designs. However, the CAT gave a much shorter test than MST under the backward assembly condition with a difference of 5 and 6 items.

For the long test (i.e., 60 items), CAT still provided shorter test length than MSTs across all conditions. The difference in averaged test length between MSTs and CATs were larger than for the test length of 45 items. CAT can have many fewer items on average to achieve similar conditional test information of MSTs, especially for the backward assembled MSTs. The CAT required shorter tests corresponding to the 1-3-3 panel design than to the 1-2-3 panel design under the forward assembly condition, but similar test length under the backward assembly condition.

In summary, different routing strategies used in MST did not affect the averaged test length in the corresponding CAT. CAT provided shorter tests than MST across all conditions on the basis of fair comparison. The difference in averaged test length between MST and CAT increased as test length increased. CAT gave much shorter tests corresponding to backward than forward assembled MST. Specifically, the 1-3-3 panel design MSTs required shorter tests in CAT than the 1-2-3 panel design MSTs did under forward assembly condition, but similar test length in CAT under backward assembly condition.

### 5.1.3 Routing Point Shift using AMI routing strategy

For the 1-2-3 panel design, the routing point from stage 2 to 3 shifted to the left quite a bit in the backward assembled MSTs. Compared to forward assembly condition, the number of examinees taking Module 3M in stage 3 under backward assembly condition was similar, while the number on Module 3E decreased and the number on Module 3H increased. For the 1-3-3 panel design under backward assembly condition, the routing point from stage 1 to 2 shifted to the left compared with the one under the forward assembly condition, and to the left further from

stage 2 to 3. This caused nearly 80 percent examinees to be routed to Module 3M and 3H evenly in stage 3. The pattern of routing point shift was the same for both test lengths.

5.1.4 MPI in the variable-length CAT

This study successfully adopted MPI to incorporate item selection algorithm and non-statistical constrains (i.e., content constrain and item exposure control) in a variable-length CAT. MSTs had equal number of items across five contents. The variable-length CAT had already matched this content requirements with great effort. There were only 1 to 2 items different across contents in CAT.

## 5.2 Discussion of Results

5.2.1 Fair comparison between MST and CAT

This study addressed a comparison of CAT and MST based on the important feature of "fairness", which is represented by different item pools for administering CAT and MST, by similar conditional standard error of measurement and by other matched properties (i.e., availability of items in item pools, maximum item exposure rate and content specifications). The results of this study offered a reference for considering testing mode change from MST to CAT for both moderate-length and long tests in terms of measurement accuracy and averaged test length in CAT.

Mean biases were similar among all MST conditions. MST and the corresponding CAT had similar mean bias.

In terms of MSE, no meaningful difference was found between MST and CAT because both procedures matched the similar conditional test information. The result of measurement accuracy implied that the efforts to make a fair comparison where successful. In addition, as indicated by Table 4.5 above, CAT provided shorter test than MST across all conditions. The difference in

averaged test length between CAT and MST increased as test length increased. However, CAT still achieved the similar measurement accuracy with MST, which confirmed the efficiency of CAT.

### 5.2.2 Testing mode change

In addition to policy issues, whether to change CAT to MST depends on measurement accuracy especially in the context of reporting continuous proficiency score, and on test length which is related to item writing cost and test administration time. Since there was no difference in measurement accuracy between routing strategies and panel designs, only the test length and assembly priority were discussed here. When considering the testing model change from CAT to MST for a moderate length test (e.g., 45 items), only 1 or 2 more items were needed for a fixed-length forward assembled MST to maintain the similar measurement accuracy. Compared with the backward assembled, CAT saved more items (i.e., 5 or 6 items). It indicated if changing CAT to a backward assembled MST, a couple of items would be added to the test which cost longer administration time and more effort in item writing. The reason was that the overall test information was lower in backward than forward assembled MST. As the term suggested, backward assembly is to assemble modules from the last to first stage following the target TIFs. For example, it assigned the most informative items from the pool to the three modules in stage 3 first, and each examinee only took one of them. The least informative items were left to the routing module in the whole panel to each examinee. In contrast, forward assembly assigned the most informative items to the three modules: 1M, 2E and 2H in a 1-2-3 MST. Each examinee took two of them, and one module in stage 3 with the least informative items in the whole panel. Figure 5.1 supports this explanation. Each subplot in Figure 5.1 is a different pathway in the 1-2-3 MST design with the test length of 45 items; different curves in a subplot represent the same

pathway in forward and backward assembled MST over 10 panels, respectively (Pathway 1 = 1M-2E-3E; Pathway 2 = 1M-2E-3M; Pathway 3 = 1M-2H-3M; Pathway 4 = 1M-2H-3H). Thus, changing CAT to the forward assembled MSTs are suggested for moderate length test. Since there is only a 1 item different between the 1-2-3 and 1-3-3 panel design, either one can be applied.

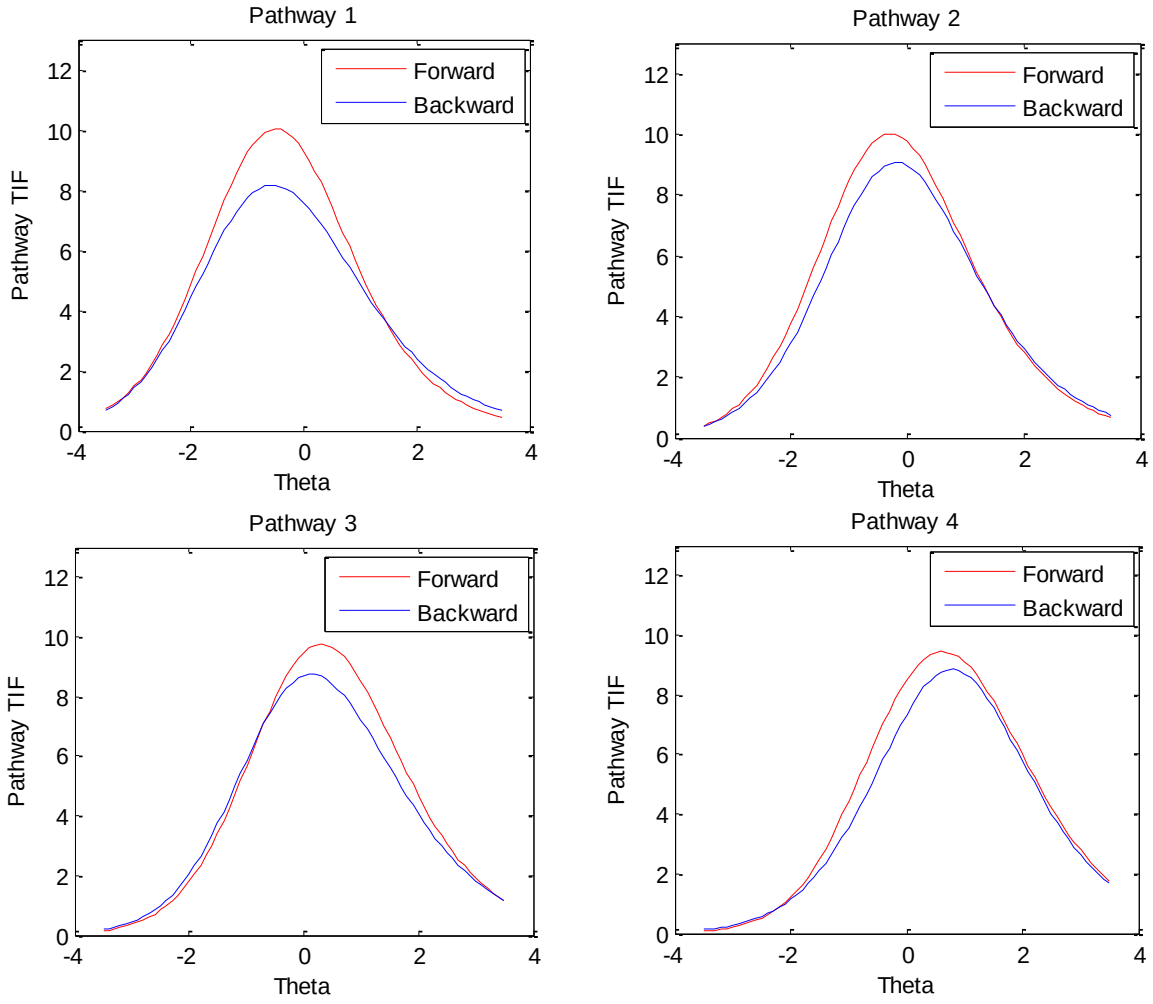


Figure 5.1 Pathway Information Curves of the 10 Parallel Panels for Both Forward and Backward Assembled 1-2-3 MST, 45 Items

When considering the testing model change from CAT to MST for a long test (e.g., 60 items), 3 to 5 more items were needed for the forward assembled MSTs to achieve similar conditional test information; and 9 to 11 more items were needed for the backward assembled MSTs. Under

such large difference in test length, CAT was still able to reach similar level of measurement accuracy with MST. When employing a three-stage MST, 3 more items were needed for the 1-2-3 panel design, while 5 or 6 more items were needed for the 1-3-3 panel design. If testing mode change is necessary, the 1-2-3 panel design was recommended.

The item-level CAT is more adaptive than MST. It assembled individualized test, and can be terminated at any point as long as the termination rule has met without losing test precision. In contrast, the module-based MST can only be terminated after a whole test form is completed. Then CAT was not surprisingly to be more efficient that MST for long tests. Generally, CAT is preferred over MST for long tests.

5.2.3 Routing point shift for using AMI routing strategy

As stated above, the routing point shifted to the left in the backward assembled MSTs, compared to the forward assembled MSTs. The reason was that the least informative items in the whole panel were assigned to stage 1 and 2 in backward assembled MSTs. As stated earlier, backward assembly assigned the most informative items from the pool to the three modules in stage 3 first, and left the least informative items to stage 1 followed by stage 2 in the whole panel. Since 3PL-IRT based model was applied in this study, the items in stage 1 and 2 have the low-discriminative ability to separate the examinees in the 1-2-3 panel design. For this reason, fewer examinees were routed to Module 3E than expected. The number of examinees on Module 3M was the largest compared with the remaining two, which was expected due to the standard normal distribution of examinee's ability level. Figure 5.2 illustrates the routing points shift between assembly priorities for the 1-2-3 panel design. The red dot represents the routing point identified by the intersection of adjacent cumulative TIFs. The same reasoning applies to the routing point shift of the 1-3-3 panel design, which figure can be found in the Appendix. In

summary, backward assembled MSTs tended to route examinees to the module whose difficulty level was higher than examinee's ability level. This brought to slightly lower measurement accuracy of backward than of forward assembled MST.

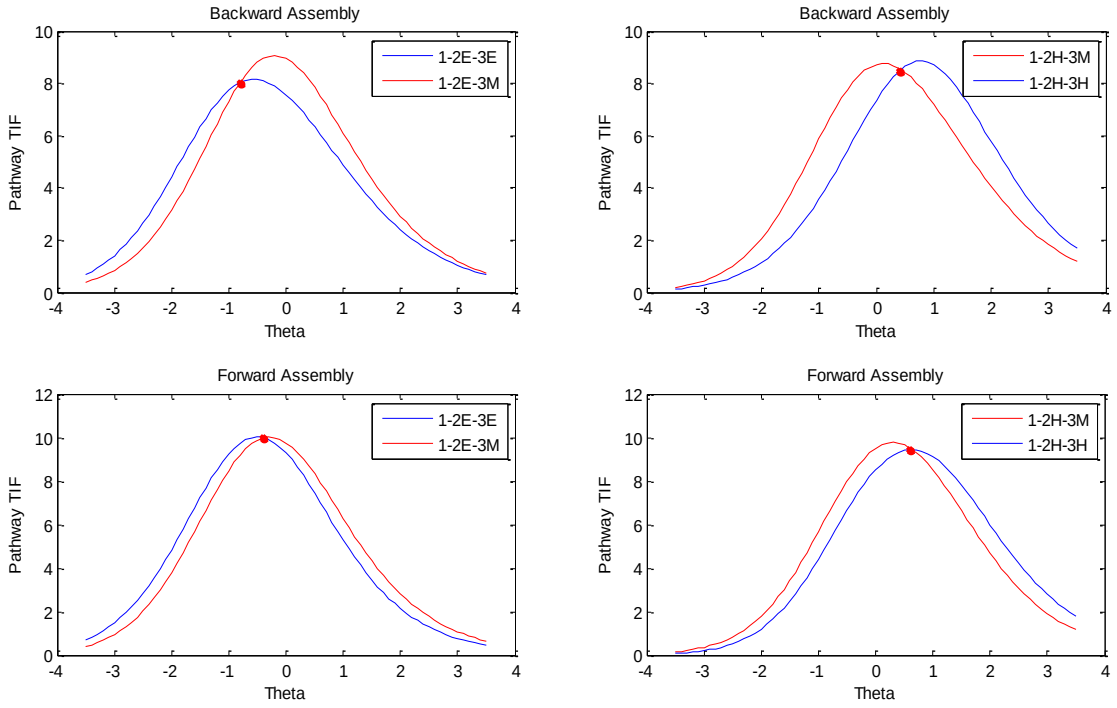


Figure 5.2 The Example of Routing Points Shift between Assembly Priorities of One of the 1-2-3 Panels

## 5.3 Implications

The major finding from this simulation study confirmed the efficiency of CAT over MST. CAT is always able to achieve similar measurement accuracy with MST by a shorter test. The fair comparison of MST and CAT provides a reference for testing mode change in terms of ability recovery and averaged test length. Previous research (e.g., Zheng, 2012) pointed out that backward assembly outperformed forward assembly in terms of classification accuracy. However, when considering testing mode change from CAT to MST, the backward assembled MST is not suggested even for a classification-oriented test. The reason was that it required a much longer test

in MST to achieve similar level of measurement accuracy of CAT. Whether to change the testing mode depends on the current averaged test length in CAT. If the current CAT has a moderate-length test, switching to a forward assembled MST with 3 stages is plausible and feasible. The routing strategies between DPI and AMI and the panel designs between 1-2-3 and 1-3-3 will not affect measurement accuracy. For a long test, staying with CAT is preferred over switching to MST.

### 5.4 Limitation and Future Studies

Although the findings of this study addressed the posed questions, they also raise other questions because of the limitations in this study. These limitations can be investigated in the future research.

First, factors to assemble the MST are limited in this study, such as test length, non-overlapping panels, equal number of items in each module, and three-stage panel design. This study applied 45 and 60 item tests because they are similar with the test lengths of many large-scale high-stakes testing program, despite its association with test reliability and decision accuracy. However, short tests using MST (e.g., NAEP) are also administered widely in reality, which should arouse the attention for future studies. Compared with non-overlapped panels, creating overlapped panels requires less items from the pool, and are supposed to satisfy the target TIF and content constraints better than non-overlapped panels. This is more practical if the operational MST pool size is limited. As mentioned earlier, mixed integer programming (MIP) as a test assembly strategy is used successfully for automated test assembly following optimization routines. It can be considered as an alternative algorithm to NWADH as for improving the ability to implement a MST. Thus, future studies can include the factors of MST varied by test assembly algorithm, panel design and see how they compared fairly with the CAT designs.

Second, all simulations adopted equal number of items for each content and each module for convenience in both item pool and administered test. In reality, content specifications always have different number of items across contents. So does an item pool. Future studies can make a comparison of CAT and MST using a real item pool followed by practical content constraints.

Third, this study assumed there were enough items to create MST item pool and panels. However, not all exams have a master pool or a large operational CAT pool. For example, the item pool size of NAEP is above 200 items, which limited the flexibility of test assembly (e.g. the number of panels, test length, and the number of different modules that an item is allowed to be assigned to). The successfulness of ATA, which depends on the quality of the item pool, is associated with the availability of items in the operational item pool. Future studies can create the MST panel with practical limitations in test assembly under a real item pool.

Fourth, this study only employ multiple choice items to create the item pool and simulated tests. Mixed format tests including multiple choice items and passages are more naturally used in a real test (e.g. AICPA). Considering both dichotomous and polytomous models may provide further information about testing mode change for a real test. Future studies can therefore examine the fair comparison of mixed-format MST and CAT.

# APPENDIX

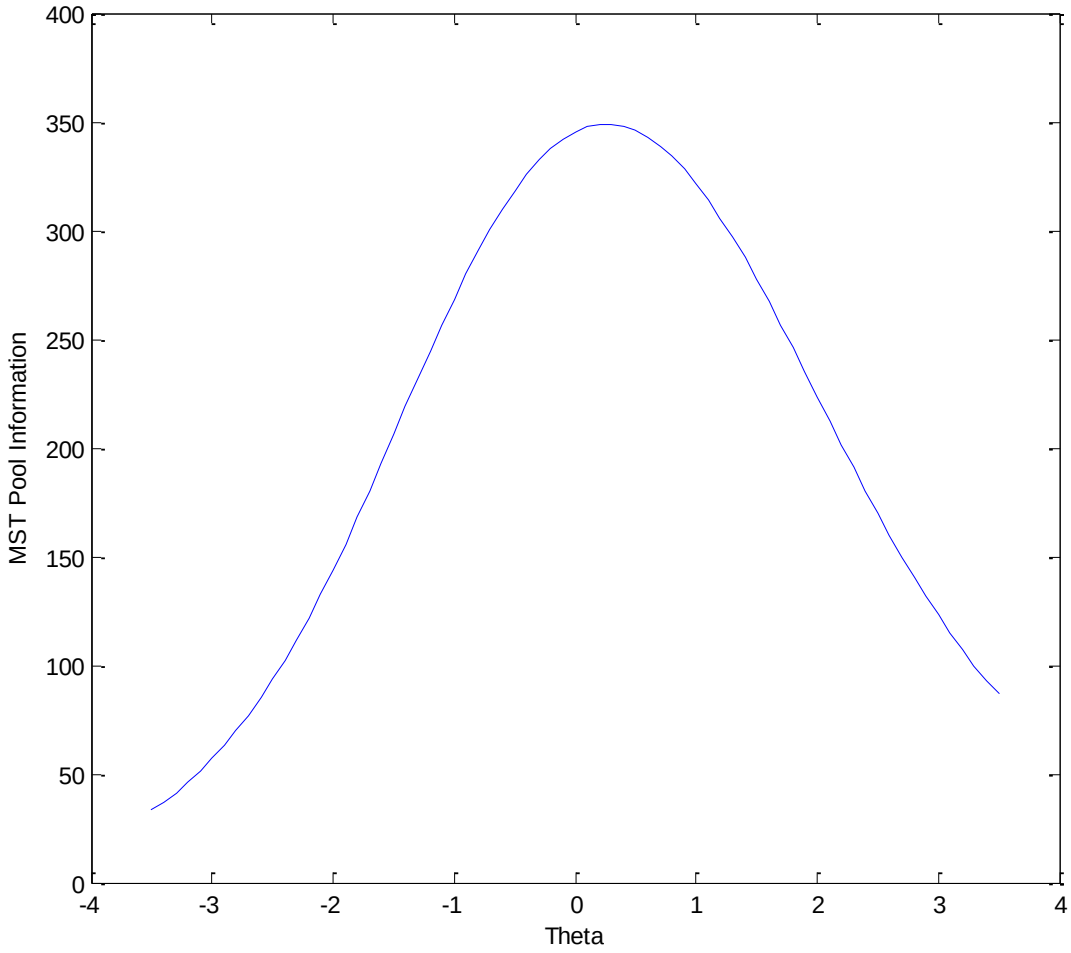


Figure A.1 MST Pool Information Curve

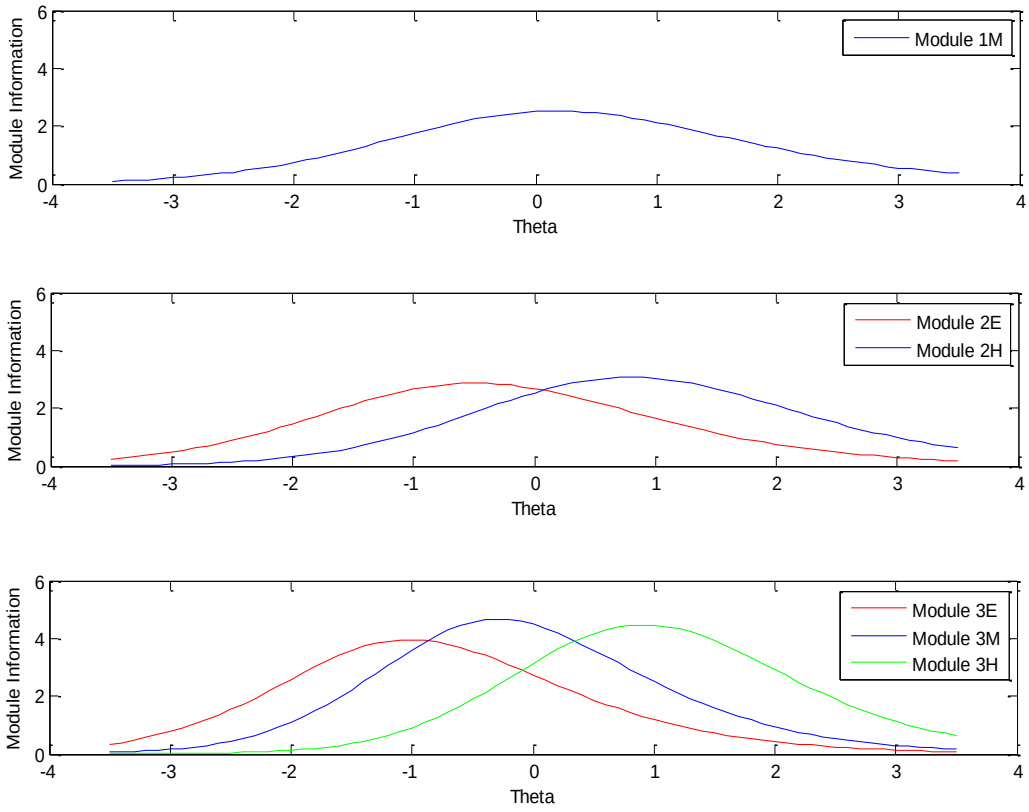


Figure A.2 Module Level Target TIFs of the 1-2-3 Panel Design, Backward Assembly, 45 Items

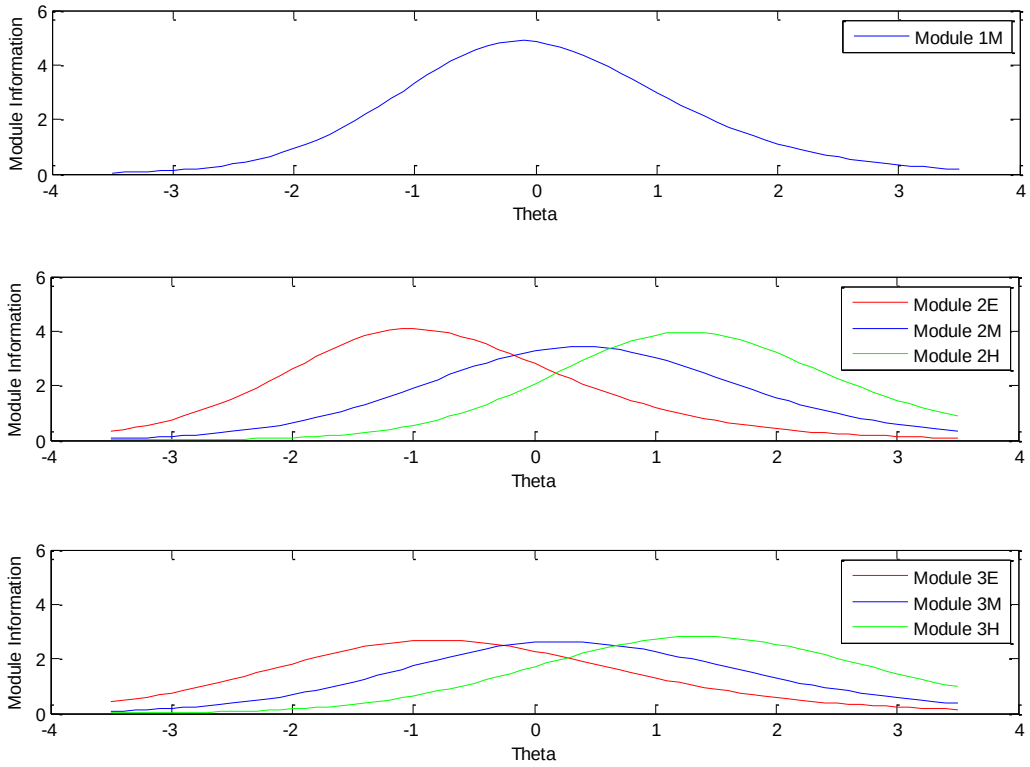


Figure A.3 Module Level Target TIFs of the 1-3-3 Panel Design, Forward Assembly, 45 Items

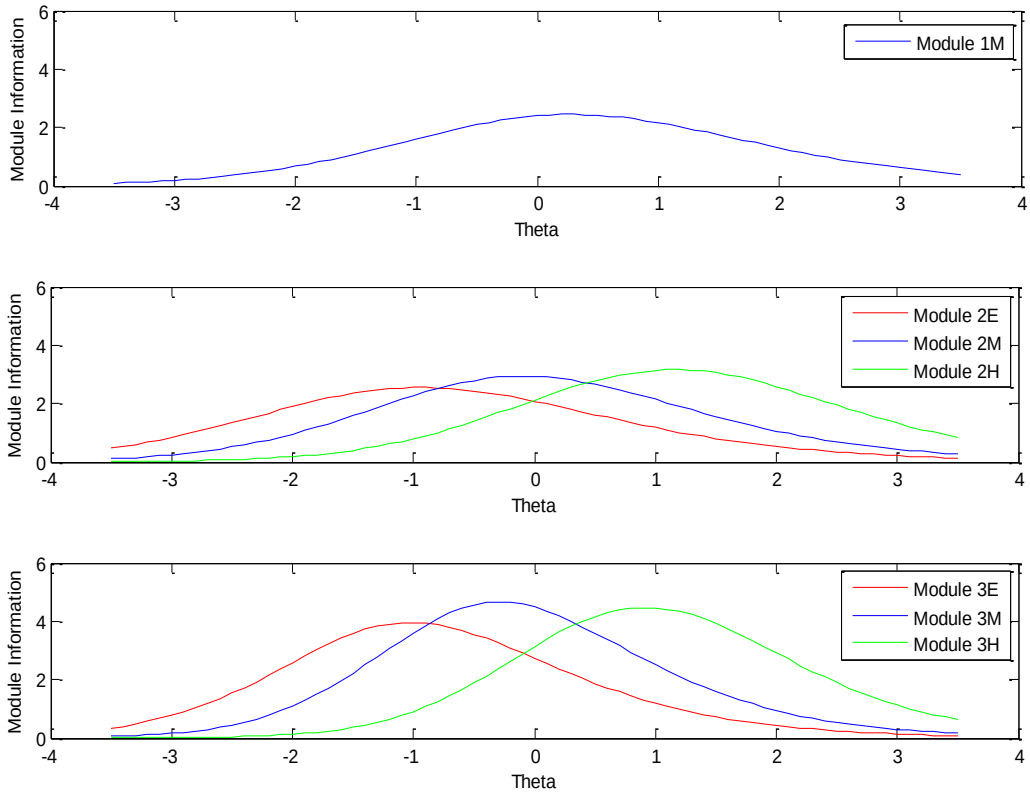


Figure A.4 Module Level Target TIFs of the 1-3-3 Panel Design, Backward Assembly, 45 Items

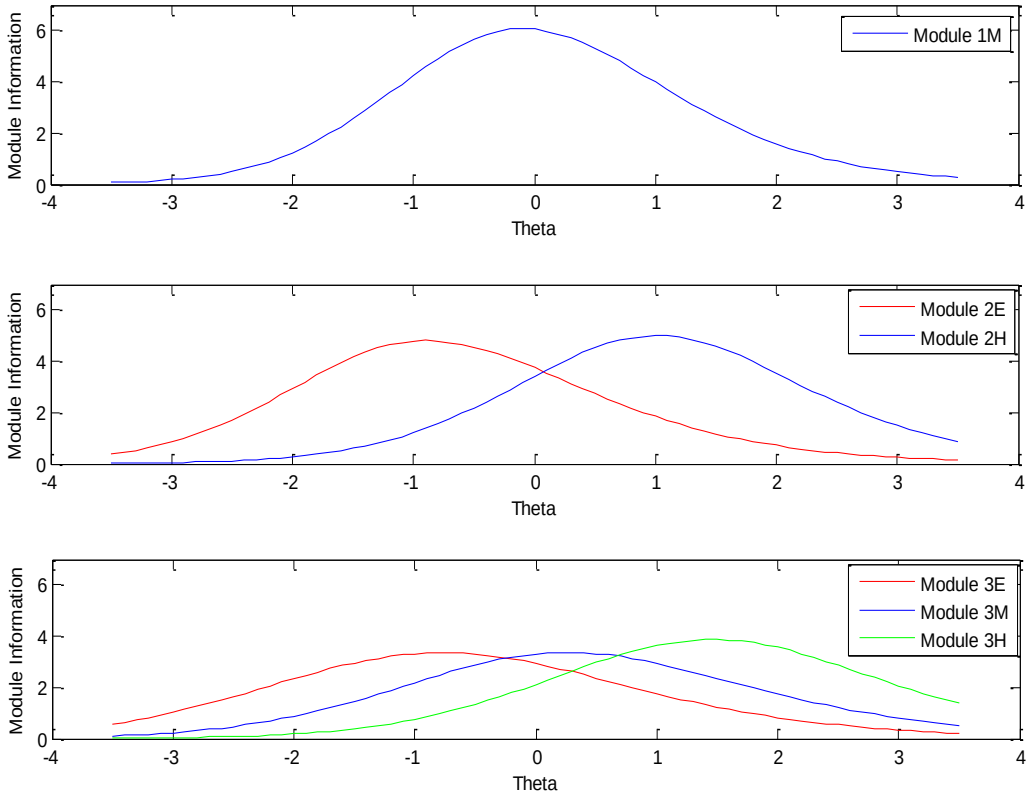


Figure A.5 Module Level Target TIFs of the 1-2-3 Panel Design, Forward Assembly, 60 Items

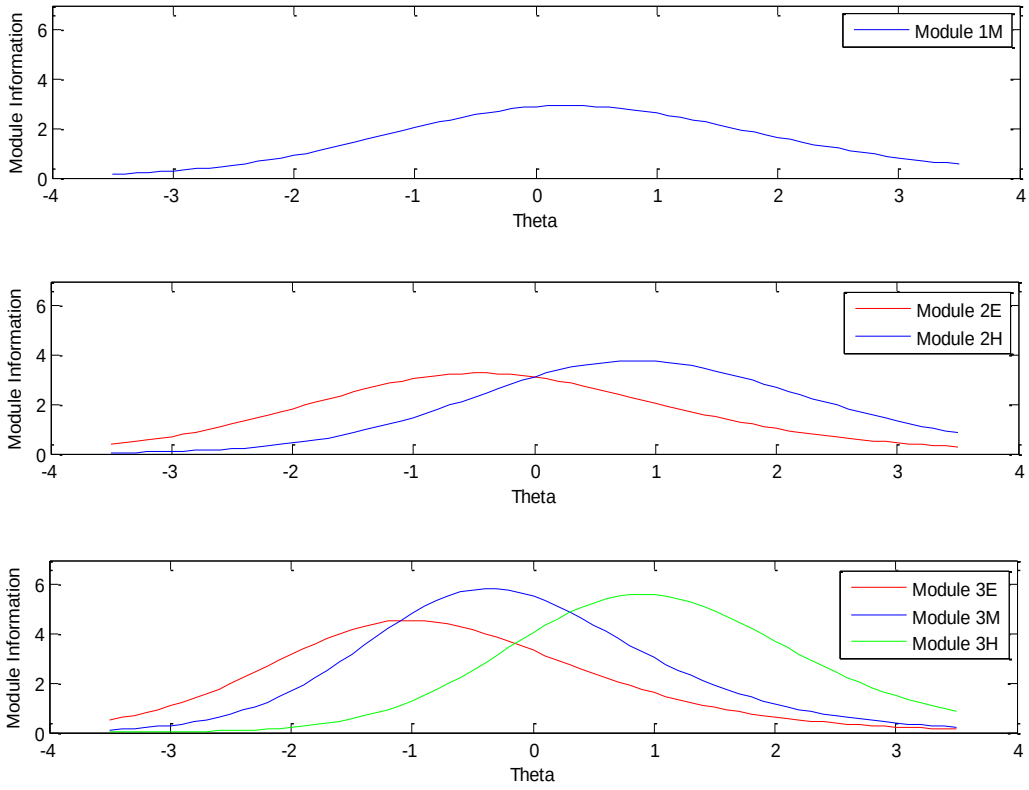


Figure A.6 Module Level Target TIFs of the 1-2-3 Panel Design, Backward Assembly, 60 Items

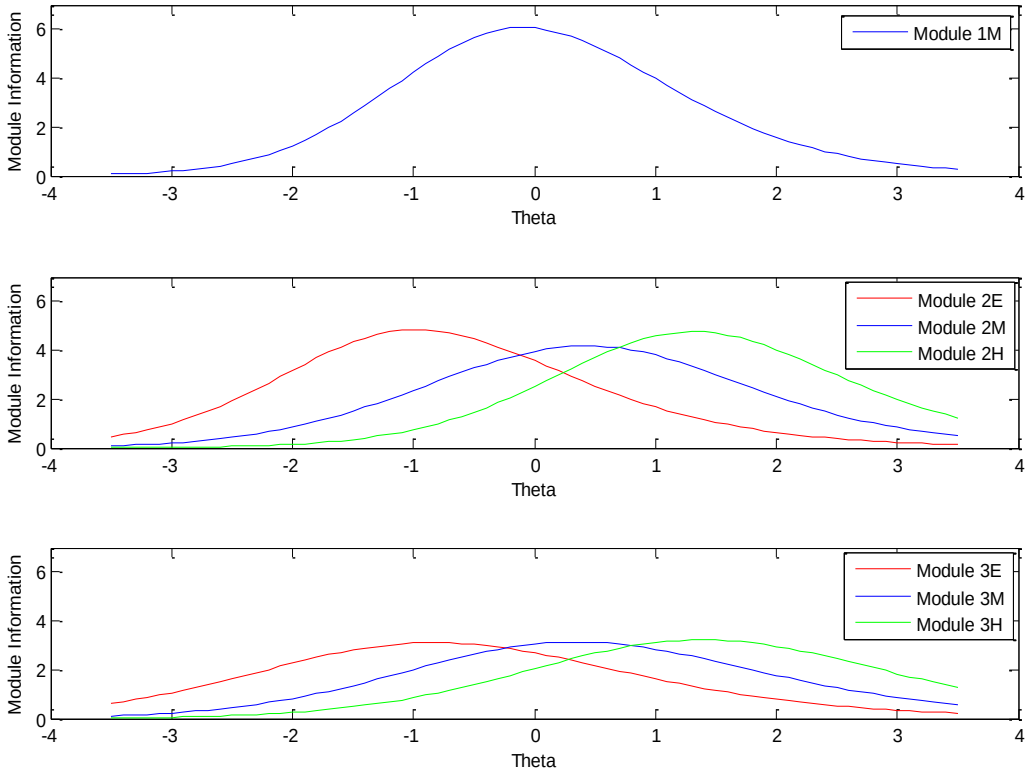


Figure A.7 Module Level Target TIFs of the 1-3-3 Panel Design, Forward Assembly, 60 Items

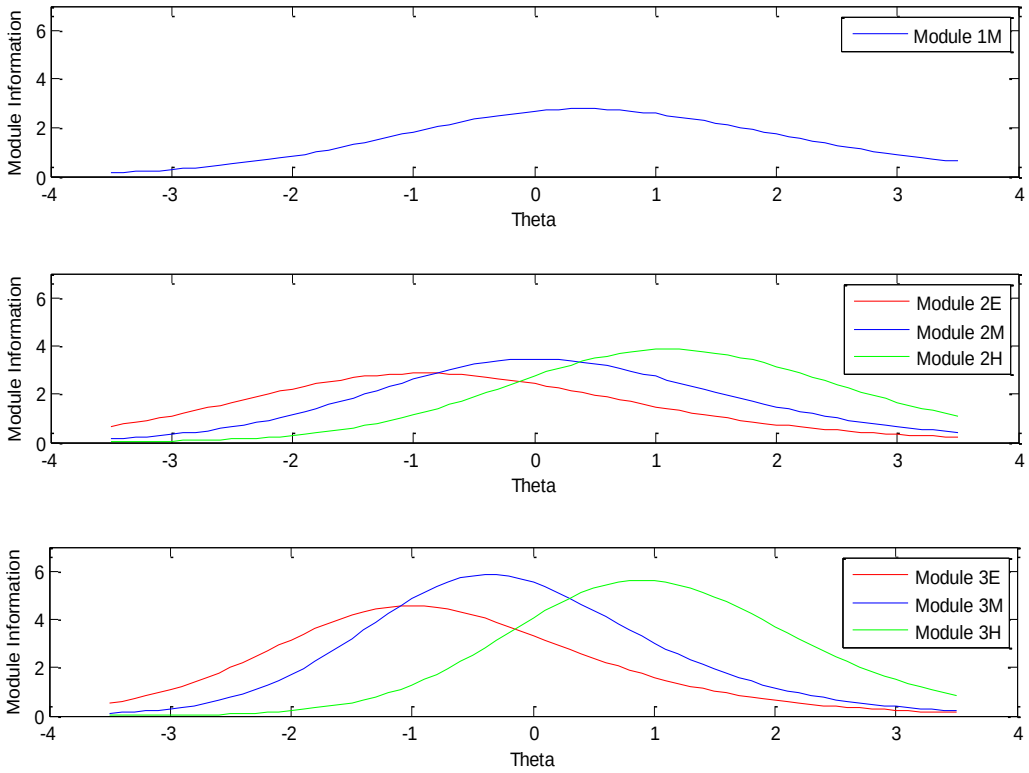


Figure A.8 Module Level Target TIFs of the 1-3-3 Panel Design, Backward Assembly, 60 Items

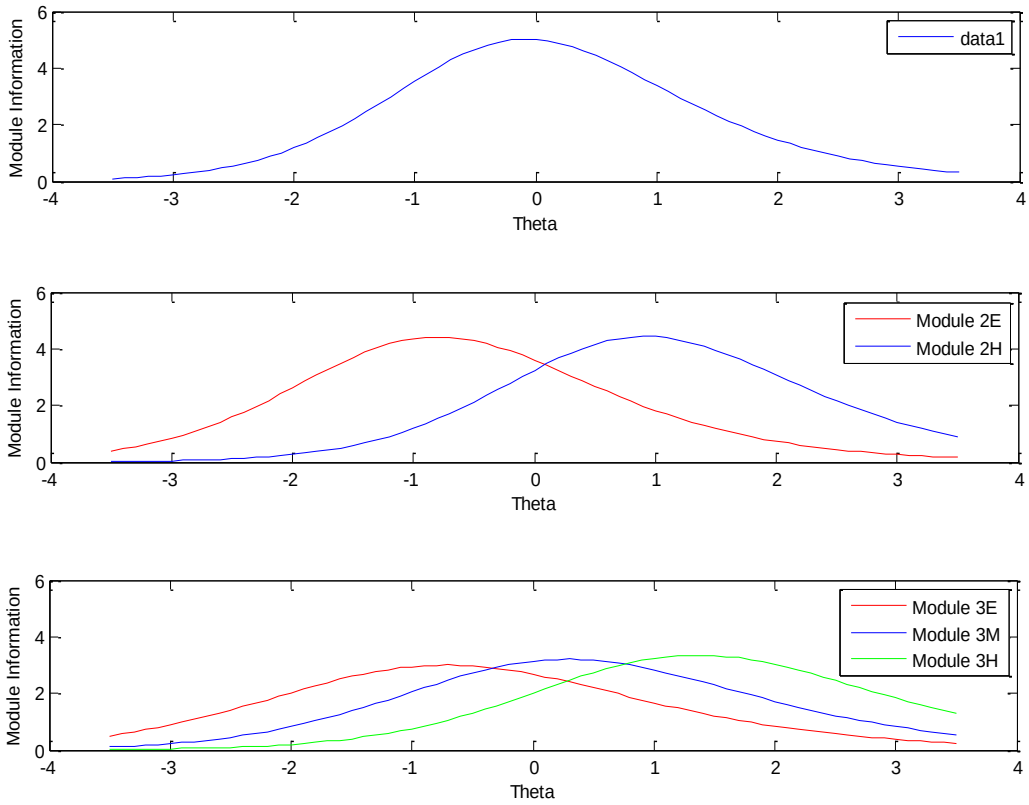


Figure A.9 Averaged Module Level Information Curves across Forward Assembled Panels for the 1-2-3 Panel Design, 60 Items

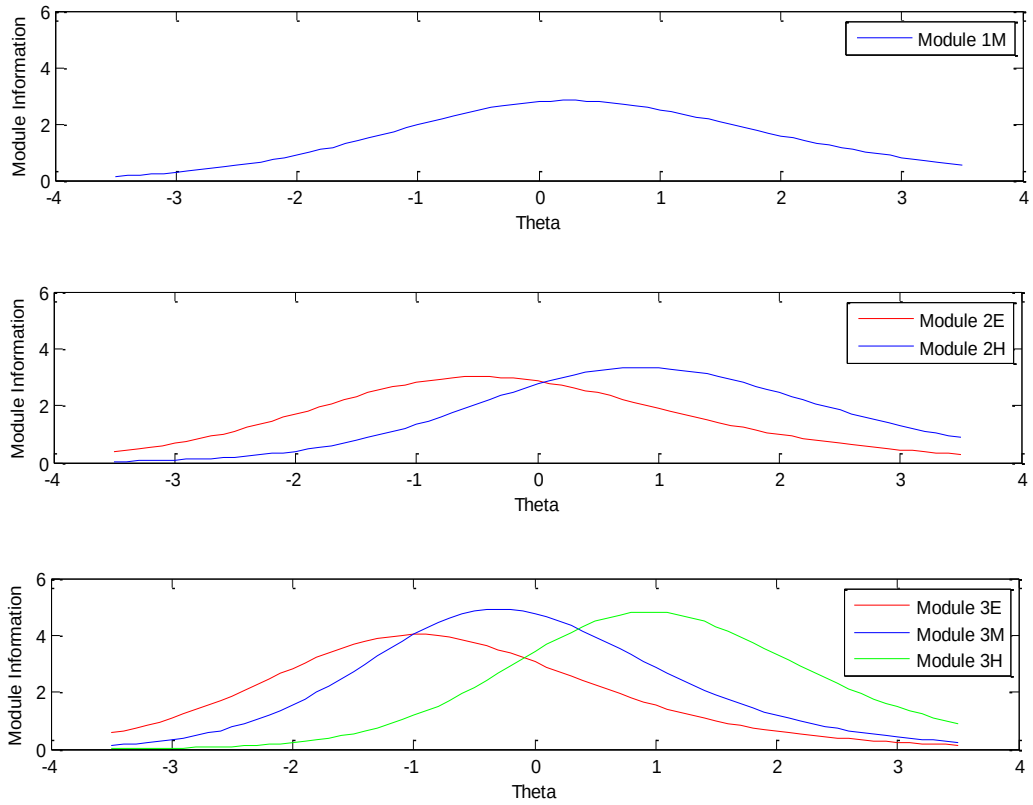


Figure A.10 Averaged Module Level Information Curves across Backward Assembled Panels for the 1-2-3 Panel Design, 60 Items

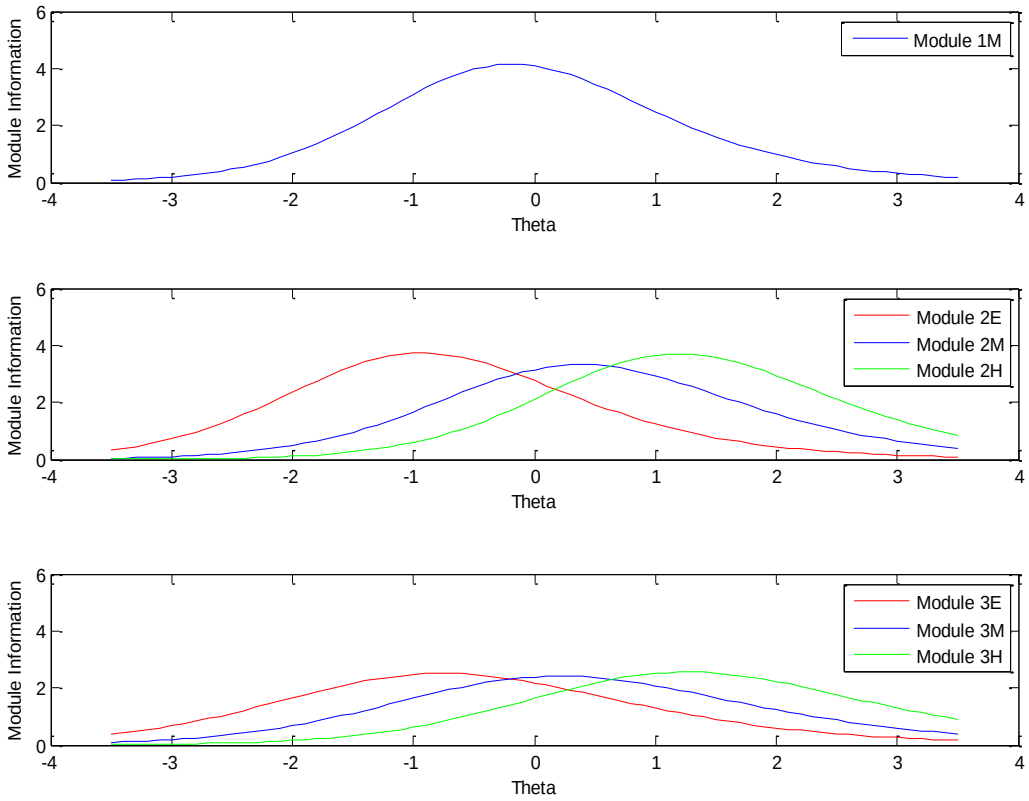


Figure A.11 Averaged Module Level Information Curves across Forward Assembled Panels for the 1-3-3 Panel Design, 45 Items

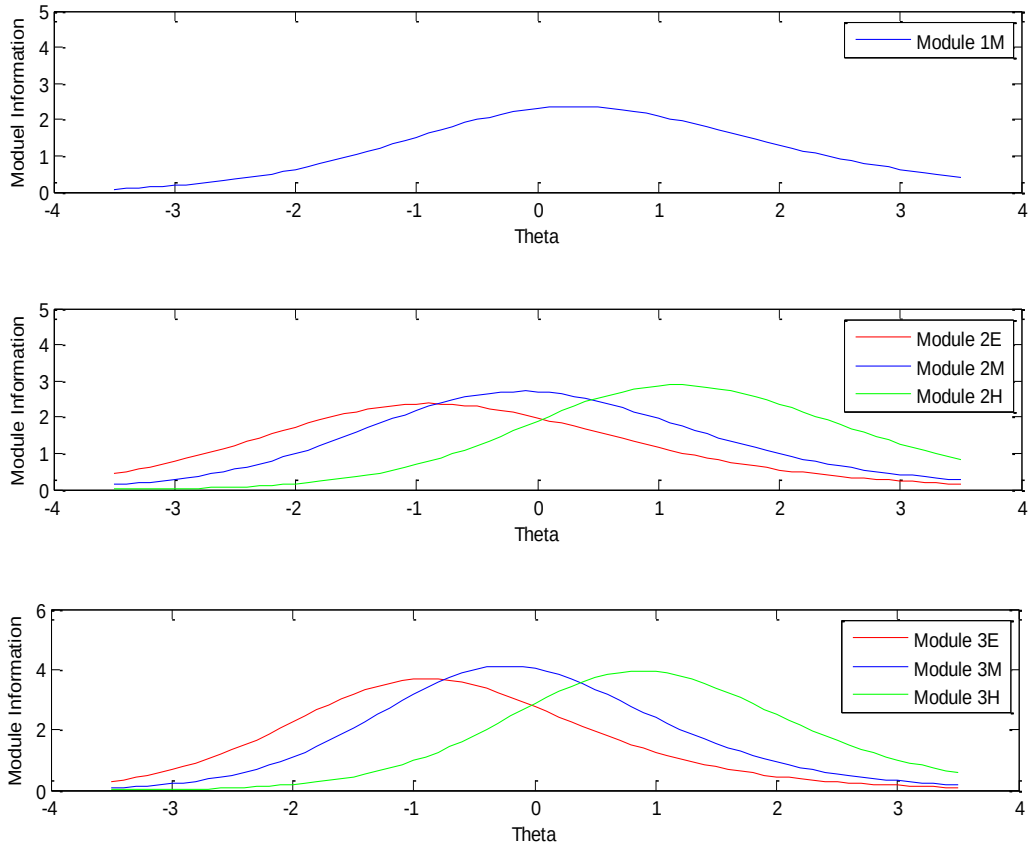


Figure A.12 Averaged Module Level Information Curves across Backward Assembled Panels for the 1-3-3 Panel Design, 45 Items

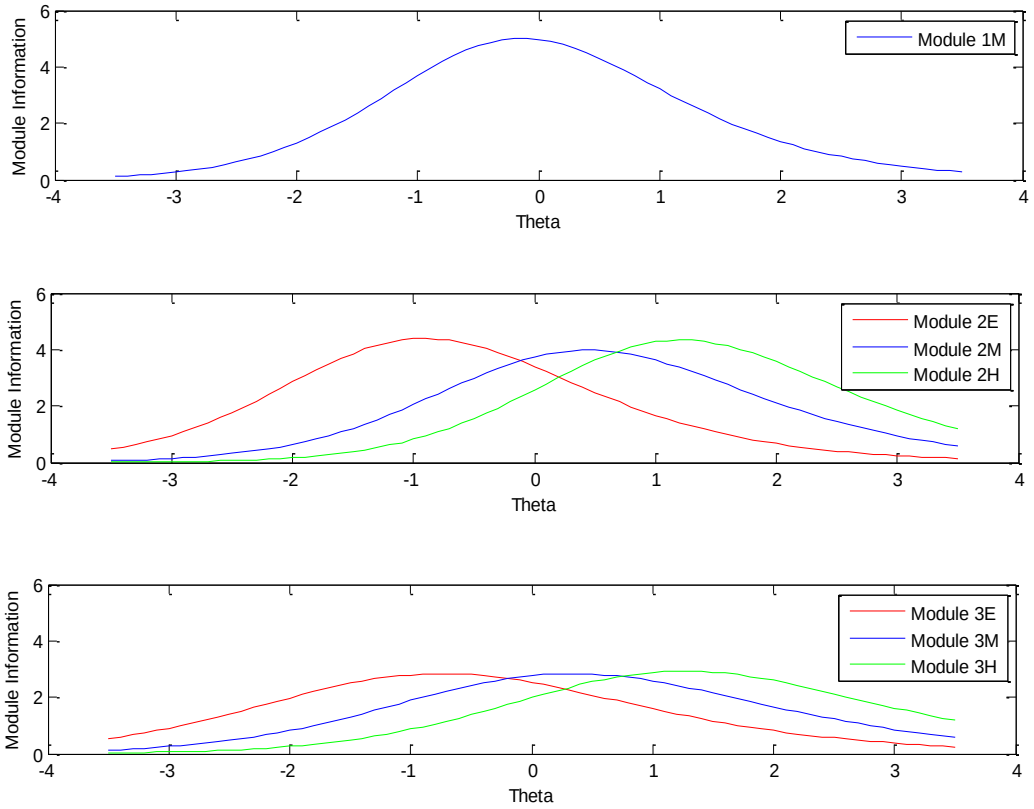


Figure A.13 Averaged Module Level Information Curves across Forward Assembled Panels for the 1-3-3 Panel Design, 60 Items

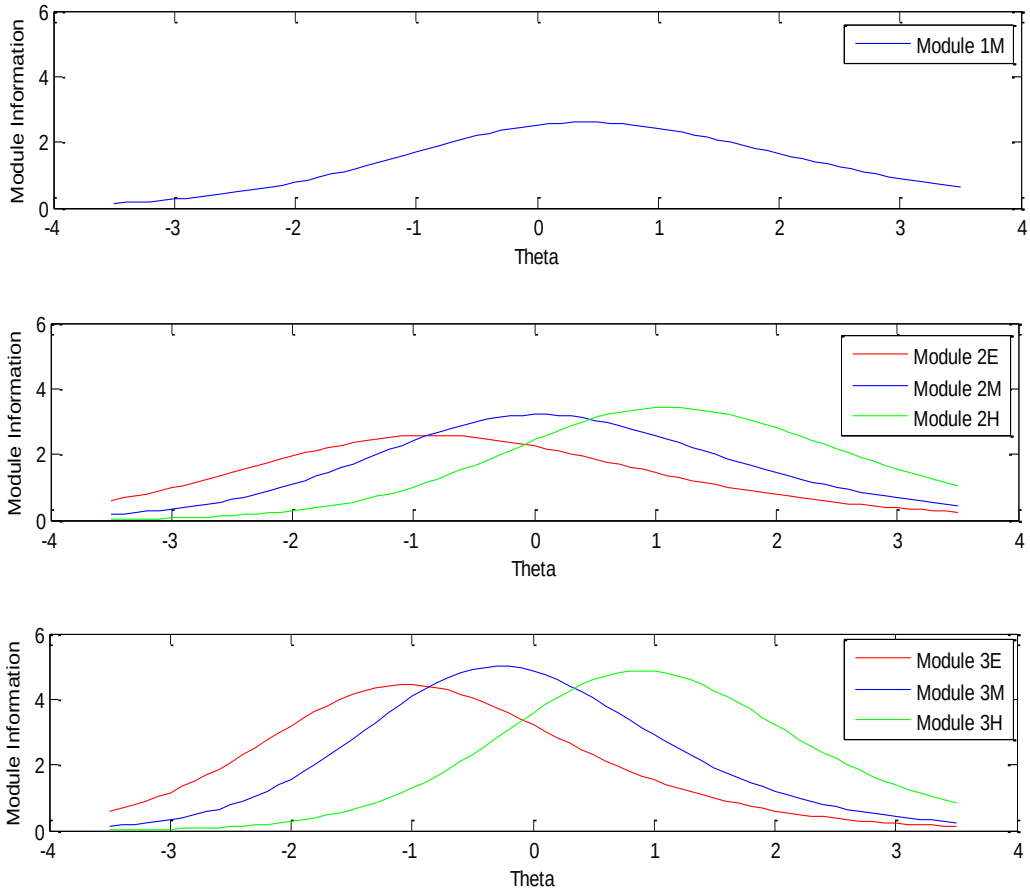


Figure A.14 Averaged Module Level Information Curves across Backward Assembled Panels for the 1-3-3 Panel Design, 60 Items

# REFERENCES

## REFERENCES


Armstrong, R. D., Jones, D. H., Koppel, N. B., & Pashley, P. J. (2004). Computerized adaptive testing with multiple-form structures. *Applied Psychological Measurement*, 28(3), 147–164.

Armstrong, R. D., & Roussos, L. (2005). *A method to determine targets for multi-stage adaptive tests*. No. 02–07). Newton, PA: Law School Admission Council.

Ariel, A., Veldkmap, Bernard P., & Breithaupt, K. (2006). Optimal testlet pool assembly for multistage testing designs. *Applied Psychological measurement*, 30, 204–215.Bergstrom, B.A., & Lunz, M. E. (1999). CAT for certification and licensure. In F. Drasgow & J.

Olson-Buchanan (Eds.), *Innovations in Computerized Assessment* (pp. 67-91). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Breithaupt, K., & Hare, D. R. (2007). Automated simultaneous assembly of multistage testlets for a high-stakes licensing examination. *Educational and Psychological Measurement*, 67(1), 5–20.

Chalhoub–Deville, M., & Deville, C. (1999). Computer Adaptive Testing in Second Language Contexts. *Annual Review of Applied Linguistics*, 19, 273–299.

Chang, H. H. (2004). Understanding computerized adaptive testing: From Robins-Monro to Lord and beyond. In David Kaplan (Ed.) *The Sage handbook of quantitative methodology for the social sciences,* (pp. 117–136). Thousand Oaks, CA: Sage Publications, Inc.

Chang, H-H., & Stout, W. (1993). The asymptotic posterior normality of the latent trait in an IRT model. *Psychometrika*, 58, 37–52.

Chang, H.-H., & Ying, Z. (1999). a-Stratified Multistage Computerized Adaptive Testing. *Applied Psychological Measurement*, 23(3), 211–222.

Chen, S., Hou, L., & Dodd, B. G. (1998). A comparison of maximum likelihood estimation and expected a posteriori estimation on CAT using the partial credit model. *Educational and Psychological Measurement*, 53, 61–77.

Cheng, Y., & Chang, H.-H. (2009). The maximum priority index method for severely constrained item selection in computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology*, 62(2), 369–383.

Davis, L. L. (2004). Strategies for controlling item exposure in computerized adaptive testing with the generalized partial credit model. *Applied Psychological Measurement*, 28(3), 165–185.

Davis, L. L., & Dodd, B. G. (2003). Item Exposure Constraints for Testlets in the Verbal Reasoning Selection of the MCAT. *Applied Psychological Measurement*, 27(5), 335–356.

Diao, Q., van der Linden., W. J. (2011a). Automated test assembly using lp_Solve version 5.5 in R. *Applied Psychological Measurement*, 35(5), 398–409.

Gu, L., & Reckase, M. (2007). Designing optimal item pools for computerized adaptive tests with Sympson-Hetter exposure control. In *Graduate Management Admission Council Conference on Computerized Adaptive Testing*, Minneapolis, MN.

Hambleton, R. K., & Xing, D. (2006). Optimal and nonoptimal computer-based test designs for making pass-fail decisions. *Applied Measurement in Education,* 19(3), 221–239.

He, Wei (2010). Optimal item pool design for a highly constrained computerized adaptive test. Unpublished doctoral dissertation, Michigan State University.

Hambleton, R. K., & Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park CA: Sage.

Hendrickson, A. (2007). An NCME instructional module on multistage testing. *Educational Measurement: Issues and Practice*, 26(2), 44–52.

Ho, T. H. (2010). *A comparison of item selection procedures using different ability estimation methods in computerized adaptive testing based on the generalized partial credit model*. University of Texas at Austin.

ILOG, Inc. (2005). CPLEX Suite (Version 9.1) [Computer software]. Mountain View, CA: Author.

Jodoin, M. G. (2003). Psychometric properties of several computer-based test designs with ideal and constrained item pools. Unpublished doctoral dissertation, University of Massachusetts at Amherst.

Jodoin, M. G., Zenisky, A., & Hambleton, R. K. (2006). Comparison of the Psychometric Properties of Several Computer-Based Test Designs for Credentialing Exams With Multiple Purposes. *Applied Measurement in Education*, 19(3), 203–220.

Park, R., Kim, J., Dodd, B. G., & Chung, H. (2011). JPLEX: Java simplex implementation with branch-and-bound search for automated test assembly. *Applied Psychological Measurement*, 35(8), 643–644.

Keng, L. (2008). A comparison of the performance of testlet-based computer adaptive tests and multistage tests. Unpublished Doctoral dissertation, The University of Texas at Austin.

Kim, H., & Plake, B. S. (1993). *Monte Carlo simulation comparison of two-stage testing and computerized adaptive testing*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Atlanta.

Kingsbury, G. G., & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive testing. *Applied Measurement in Education*, 2, 359–375.

Kingsbury, G. G., & Zara, A. R. (1991). A comparison of procedures for content-sensitive item selection in computerized adaptive tests. *Applied Measurement in Education*, 4(3), 241–261.

Leung, C. K., Chang, H. H., & Hau, K. T. (2005). Computerized adaptive testing: A mixture item selection approach for constrained situations. *British Journal of Mathematical and Statistical Psychology*, *58*, 239–257.

Lord, F. M. (1971). A theoretical study of two-stage testing. *Psychometrika*, *36*, 227–242.

Lord, F. M. (1971). The self-scoring flexilevel test. *Journal of Educational Measurement*, *8*, 147–151.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale NJ: Erlbaum.

Luecht, R. (2000). Implementing the CAST framework to mass produce high quality computer-adaptive and mastery tests. In *Annual Meeting of National Council on Measurement in Education*. New Orleans, LA.

Luecht, R. M., Brumfield, T., & Breithaupt, K. (2006). A testlet assembly design for adaptive multistage tests. *Applied Measurement in Education*, 19(3), 189–202.

Luecht, R. M., & Nungester, R. J. (1998). Some practical examples of computer-adaptive sequential testing. Journal of Educational Measurement, 35, 229–249.

Luecht, R. M., Nungester, R. J., & Hadidi, A. (1996). Heuristic-based CAT: Balancing item information, content and exposure. In *Annual Meeting of the National Council on Measurement in Education*, New York.

Luecht, R., & Burgin, W. (2003). Test Information Targeting Strategies for Adaptive Multistage Testing Designs. In *Annual Meeting of National Council on Measurement in Education*. Chicago, IL.

Luecht, R. (1998). Computer-Assisted Test Assembly Using Optimization Heuristics. *Applied Psychological Measurement*, 22(3), 224–236.

Lunz, M. E., Bergstrom, B. A., & Wright, B. D. (1992). The effect of review on student ability and test efficiency for computerized adaptive tests. *Applied Psychological Measurement*, 16(1), 33–40.

Mao, L. (2014). *Designing p-Optimal Item Pools for Multidimensional Computerized Adaptive Testing*. Unpublished Doctoral Dissertation. Michigan State University.

McBride, J. R., & Martin, J. T. (1983). Reliability and validity of adaptive ability tests in a military setting. In D. J. Weiss (Ed.), *New horizon in testing* (pp. 223–226). New York: Adacemic Press.

Melican, G. J., Breithaupt, K., & Zhang, Y. (2010). Designing and implementing a multisatage adaptive test: The uniform CPA exam. In W. J. van der Linden & C. a. W. Glas (Eds.), *Elements of adaptive testing* (pp. 167–190). New York, NY: Springer.

Owen, R. J. (1975). A Bayesian Sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association*, 70, 351–356.

Patsula, L. N. (1999). *Comparison of computerized adaptive testing and multi-stage testing*. Unpublished doctoral dissertation, University of Massachusetts at Amherst.

Patsula, L. N. & Steffan, M. (1997). *Maintaing item and test security in a CAT environment: A simulation study*. Paper presented at the Annual Meeting of the National Council on Measurement in Education.

Reckase, M. (1989). Adaptive testing: The evolution of a good idea. *Educational Measurement Issues and Practice*, 8(3), 11–15.

Reckase, M. (2007). The design of p-optimal item bank for computerized adaptive tests. In D. J. Weiss (Ed.). *Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing*. Retrieved from www.psych.umn.edu/psylabs/CATCentral/

Spray, J. A. & Reckase, M. D., (1994, April). *The selection of test items for decision making with a computer adaptive test*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Stark, S., & Chernyshenko, O. S. (2006). Multistage testing; widely or narrowly applicable? *Applied Measurement in Education*, 19(3), 257–260.

Stocking, M. L., & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement*, 17(3), 277–292.

Sympson, J. B., & Hetter, R. D. (1985). *Controlling item-exposure rates in computerized adaptive testing*. Paper presented at the 27th Annual Meeting of the Military Testing Association, San Diego, CA.

The MathWorks Inc. (2011). MATLAB version 7.12.0. Natick, Massachusetts: The MathWorks Inc.

The MathWorks Inc. (2015b). MATLAB version 10.1. Natick, Massachusetts: The MathWorks Inc.

The MathWorks Inc. (2016a). MATLAB version 9.0. Natick, Massachusetts: The MathWorks Inc.

Thissen, D., & Mislevy, R. J. (2000). Testing algorithms. In H. Wainer (Ed.), *Computerized adaptive testing: a primer* (pp. 101-133). Mahwah, NJ: Lawrence Erlbaum Associates.

Van der Linden, W. J. (2005). *Linear Models for optimal test design*. New York: Springer-Verlag.

Van der Linden, W. J., Ariel, A. & Veldkamp, Bernard P. (2006, Spring). Assembling a computerized adaptive testing item pool as a set of linear tests. *Journal of Educational and Behavioral Statistics*, 31(1), 81–100.

Van der Linden, W. J. & Reese, L. (1998). A model for optimal constrained adaptive testing. *Applied Psychological Measurement*, 22(3), 259–270.

Veldkamp, B. P. (2003). Item selection in polytomous CAT. In *New developments in psychometrics* (pp. 207–214). Springer Japan.

Wainer, H. (1990). Computerized adaptive testing: A prime. In. Hillsdale, NJ: Lawrence Erlbaum Associates.

Wang, X., Fluegge, L., & Luecht, R. (2012). A Large-scale Comparative Study of the Accuracy and Efficiency of ca-MST Panel Design Configurations. In *Annual Meeting of the National Council on Measurement in Education*. Vancouver, British Columbia, Canada.

Wang, T., & Vispoel, W. P. (1998). Properties of ability estimation methods in computerized adaptive testing. *Journal of Educational Measurement*, *35*(2), 109–135.

Wang, X. (2013). An Investigation on Computer-Adaptive Multistage Testing Panels for Multidimensional Assessment. Unpublished Doctoral Dissertation. The University of North Carolina at Greensboro.

Wang, S., & Wang, T. (2001). Precision of Warm's weighted likelihood estimates for a polytomous model in computerized adaptive testing. *Applied Psychological*. *Measurement*, *25*(4), 317–331.

Way, W. D. (1998). Protecting the integrity of computerized testing item pools. *Educational Measurement: Issues and Practice*, 17, 17–27.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement, 6,* 473–492.

Weiss, D. J. (1985). Adaptive testing by computer. *Journal of Consulting and Clinical Psychology*, 53, 774–789.

Zenisky, A. (2004). *Evaluating the effects of several multi-stage testing design variables on selected psychometric outcomes for certification and licensure assessment* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3136800)

Zenisky, A., Hambleton, R., & Luecht, R. (2010). Multistage Testing: Issues, Designs, and Research. In van der Linden, W. & Glas, C. (Eds.), Elements of Adaptive Testing (pp. 355–372). *New York, NY: Springer New York.*

Zenisky, A., & Hambleton, R. (2014). Multistage test designs: moving research results into practice. In Yan, D., Von Davier, A., & Lewis, C. (Eds.), *Computerized Multistage Testing* (pp. 21–36). CRC Press.

Zheng, Y., Nozawa, Y., Gao, X., & Chang, H. (2012). Multistage Adaptive Testing for a Large-Scale Classification Test: The Designs, Heuristic Assembly, and Comparison with Other Testing Modes. In *Annual Meeting of the National Council on Measurement in Education*. Vancouver, British Columbia, Canada.