ENHANCING ITEM POOL UTILIZATION WHEN DESIGNING MULTISTAGE COMPUTERIZED ADAPTIVE TESTS

By

Lihong Yang

A DISSERTATION

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

Measurement and Quantitative Methods-Doctor of Philosophy

2016

ABSTRACT

ENHANCING ITEM POOL UTILIZATION WHEN DESIGNING MULTISTAGE COMPUTERIZED ADAPTIVE TESTS

By

Lihong Yang

In recent years, the multistage adaptive test (MST) has gained increasing popularity in the field of educational measurement and operational testing. MST refers to a test in which pre-constructed sets of items are administered adaptively and are scored as a unit (Hendrickson, 2007). As a special case of Computerized Adaptive Testing (CAT), a MST program needs the following components: an item response theory (IRT) model or non-IRT-based alternatives; an item pool design; module assembly; ability estimation; routing algorithm; and scoring (Yan et al., 2014). A significant amount of research has been conducted on components like module assembly, ability estimation, routing and scoring, but few studies have addressed the component of item pool design. An item pool is defined as consisting of a maximal number of combinations of items that meet all content specifications for a test and provide sufficient item information for estimation at a series of ability levels (van der Linden et al., 2006). An item pool design is very important because any successful MST assembly is inseparable from an optimal item pool that provides sufficient and high-quality items (Luecht & Nungester, 1998).

Reckase (2003, 2010) developed the *p*-optimality method to design optimal item pools using the unidimensional Rasch model in CAT, and it has been proved to be efficient for different item types and IRT models. The present study extended this method to MST context in supporting and developing different MST panel designs under different test configurations. The study compared the performance of the MST assembled under the most popularly studied panel designs in the literature, such as 1-2, 1-3, 1-2-2, and 1-2-3. A combination of short, medium and

long tests with different routing test proportions were used to build up different tests. Using one of the most popularly investigated IRT models, the Rasch model, simulated optimal item pools were generated with and without practical constraints of exposure control. A total number of 72 optimal items pools were generated and the measurement accuracy was evaluated by an overall sample and conditional sample using various statistical measures. The *p*-optimality method was also applied in an operational MST licensure test to see if it is feasible in supporting test assembly and achieving sufficient measurement accuracy in practice.

Results showed that the different MST panel designs achieved sufficient measurement accuracy by using the items from the optimal item pools built with the *p*-optimality method. The same was true with the operational item pool. Measurement accuracy was related to test length, but not so much to the routing test proportions. Exposure control affected the item pool size, but the distributions of the item parameters and item pool characteristics for all the MST panel designs were similar under the two conditions. The item pool sizes under the exposure control conditions were several times larger than those under no exposure control, depending on the types of MST panel designs and routing test proportions. The results from this study provide information for how to enhance item pool utilization when designing multistage computerized adaptive tests, facilitating the MST assembly process, and improving the scoring accuracy.

ACKNOWLEDGEMENTS

I am deeply indebted to my academic advisor and dissertation chair, Dr. Mark D. Reckase, who guided me through every phase of my dissertation research as well as my doctoral study. Dr. Reckase always encouraged me and supported my academic endeavors whenever needed, and guided and inspired me with his profound knowledge and professional skills along the way. I would not have completed this dissertation without his sustained support and valuable guidance. I would also like to express my sincere appreciation to the other members of my dissertation committee: Dr. Richard Houang, Dr. Joseph Martineau and Dr. Ryan Bowles, all of whom have not only contributed valuable and inspiring suggestions to my dissertation study, but also demonstrated continued help and understanding during the entire process.

My sincere thanks go to the other faculty members in Measurement and Quantitative Methods program (MQM) at Michigan State University (MSU): Dr. Kenneth Frank, Dr. Spyros Konstantopoulos, Dr. Kim Maier, Dr. Tenko Raykov, and Dr. William Schmidt, who have taught me with enlightening lectures on quantitative methods and psychometrics which laid a solid foundation for my future research.

I would like to thank the MQM program for providing me with both the research assistantship and teaching assistantship opportunities at the College of Education; and thank Dr. Douglas Estry, and Dr. Karen P. Williams for offering me the two great research assistantship opportunities which allow me to work for both the large-scale longitudinal study and large-scale student assessment project, and gain valuable experiences in real data analysis and research. My special acknowledgement goes to my two research supervisors Dr. Cristian Meghea and Dr. Julie Libarkin for their superb and valuable guidance to my research, and

persistent care to my study as well as my personal life. I am also grateful to Dr. Guofang Li for all her valuable instructions and inspirations to me in the several research projects that we have cooperated.

My sincere gratitude also goes to Houghton Mifflin Harcourt which offered me a great psychometric internship opportunity in 2014 and invited me for a second productive internship in 2015. I gained many valuable hands-on experiences in operational testing while working on the various research projects using their large-scale student assessment data. I hope to acknowledge my thanks to Dr. Doug Becker, Dr. Stephen Murphy, Dr. Diane Signatur, Dr. John Denbleyker, Dr. Rong Jin, and all the research and measurement team for their generous support to me during the two internships.

I wish to thank my friends and colleagues, Minh Duong, Anne Traynor, Cheng-Hsien Li, Emre Gonulates, Liyang Mao, Hong Qian, Bing Tong, Xin Luo, Wei Li, Changhui Zhang, Xuechun Zhou, Hyesuk Jang, and Ifeoma Iyioke for their warm support to my study and life at Michigan State University.

Last but not the least, I would like to express my heartfelt thanks to my parents, my parents-in-law, my husband and my children for their unconditional love and support to me in the process of my doctoral study, which gave me endless strength to move forward.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: Introduction

In recent years, the multistage adaptive testing (MST) has gained increasing popularity in the field of educational measurement and operational testing. MST refers to a test in which pre-constructed sets of items are administered adaptively and are scored as a unit (Hendrickson, 2007). More specifically, an MST instrument generally begins with a first-stage module or routing test with medium item difficulty. Examinees are adaptively routed to advanced-stage modules based on their performance on the routing test. A complete route an examinee takes in the MST is termed a "pathway" and the assembled MST is a "panel". A panel usually has several different pathways with two or three stages and two or three modules at each of the advanced stages. (Luecht & Nungester, 1998).

A sample MST panel design by Luecht et al. (1998) is shown in the following figure:

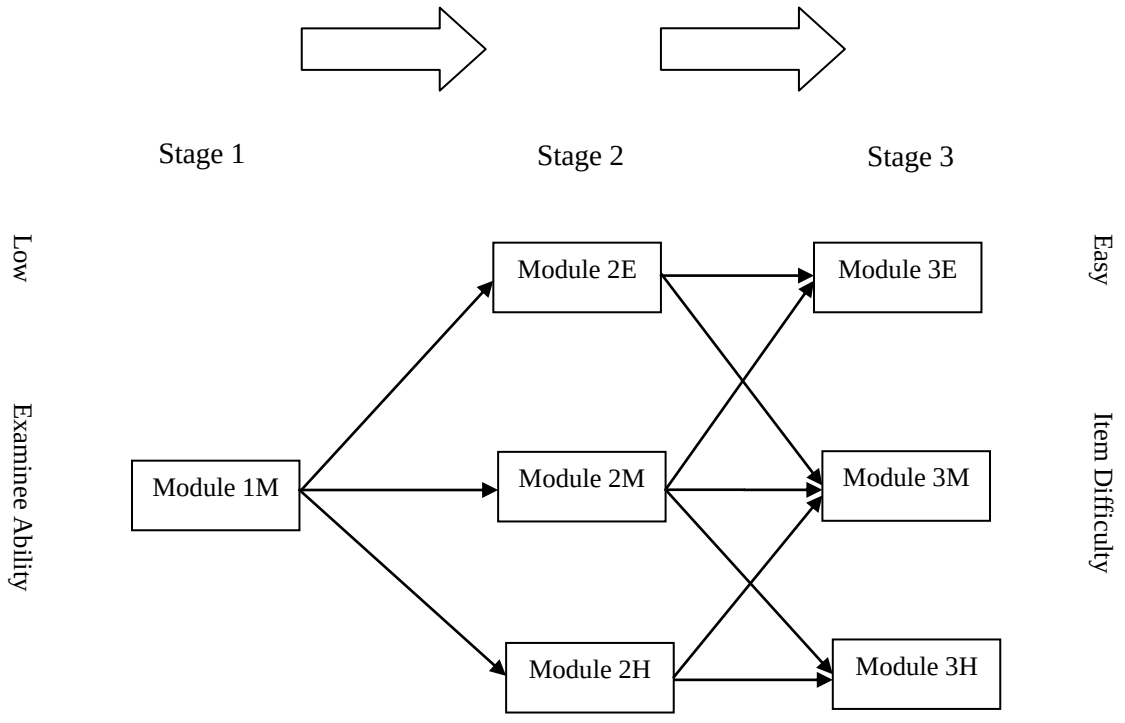

Figure 1.1 A conceptual description of a 1-3-3 MST panel design (Luecht & Nungester, 1998)
Note: E means Easy, M means Moderate and H means Hard in terms of the average item difficulty range for the module

This MST panel has one routing test at Stage 1, three modules at Stage 2 and three modules at Stage 3. During the test delivery, one module from each stage is administered to each examinee. The procedure of this MST administration is: 1) an examinee is administered Module 1M, 2) based on the estimated ability, the examinee is routed to one of the modules at Stage 2, 3) the estimated ability from Stage 2 is used to route the examinee to one of the modules at Stage 3. An examinee with low proficiency might take pathway Module 1M→Module 2E→Module 3E or Module 1M→Module 2E→Module 3M, and an examinee with high proficiency might see Module 1M→Module 2H→Module 3H or Module 1M→Module 2H→Module 3M. Each examinee is only assigned to take one pathway and the modules are administered sequentially and adaptively within panel across stages.

One exceptional feature of the MST is that test forms are pre-constructed before administration so that the equivalence of psychometric properties and content coverage for the test can be ensured. With MST, adaptation occurs at the module level instead of the item level as in Computerized Adaptive Testing (CAT). This results in fewer adaptation points, more efficient test assembly, and well-controlled content-balancing (Berger, 1994; Luecht, 2000). In CAT, due to the fact that item selection favors the most informative item for the provisional ability estimate, the probability for examinees with similar abilities to receive the same set of items is high. Consequently, the test scores obtained are faced with validity threat caused by possible item pre-knowledge (Patsula & Hambleton, 1999). Comparatively speaking, with MST, since the test is pre-constructed, the item and test exposure risk could be addressed prior to test administrations (Hendrickson, 2007). MST is also a compromise between CAT and Paper & Pencil (P&P) testing. Like CAT, MST is characterized by easier data collection, faster score reporting, easier control of test standardization and test security, shorter administration time, and

the possibility of applying innovative item types (Chalhoub-Deville & Deville, 1999, Rotou et al., 2003). Similar to P&P testing, examinees in MST are allowed to review their answers within item sets. This reduces the likelihood of chance errors that are caused by negligence or time restrictions in their initial attempt with the test items.

Due to the obvious benefits, many testing programs have successfully shifted to MST over the past decade. For example, in 2004, the Certified Public Accountants (AICPA) examination was switched from P&P to MST (Luecht, Brumfield, & Breithaupt, 2006). In August, 2011, the Graduate Record Examination (GRE) also used MST to replace P&P and CAT (Zheng et al., 2012). Examples of MSTs are also found in many large-scale international educational assessment tests (e.g., The Program for the International Assessment of Adult Competencies, Chen, Yamamoto & von Davier, 2014) and K-12 testing programs (e.g., the Educational Records Bureau Comprehensive Testing Program, Wentzel, Mills, & Meara, 2014).

As summarized in Yan, Lewis, & von Davier (2014), as a special case of CAT, an MST program needs the following components: 1) an item response theory (IRT) model or non-IRT-based alternatives (e.g., tree-based methodology for MST); 2) an item pool design; 3) module assembly; 4) ability estimation; 5) routing algorithm; and 6) scoring. Among the six MST components, an item pool design is very important because any successful MST assembly is inseparable from an optimal item pool that provides sufficient and high-quality items (Luecht & Nungester, 1998). An item pool is defined as consisting of a maximal number of combinations of items that meet all content specifications for a test and provide sufficient item information for estimation at a series of ability levels (van der Linden, Ariel, and Veldkamp, 2006). An item pool design is different from item pool assembly (e.g., an item pool is assembled from a master pool based on the desired test specifications). The item pool design focuses on developing an item

pool blueprint in which the distribution of numbers of items with all relevant statistical and non-statistical attributes of the items are described. The technical aspects deal with such issues as item selection algorithm, exposure control, stopping rules, and item overlap restriction, and the non-technical aspects deal with issues as target examinee population and test purpose (He & Reckase, 2014).

Two approaches have emerged for CAT item pool design: the integer programming approach developed by Veldkamp & van der Linden (2000), and the heuristic approach, as represented by the *p*-optimality method or *r*-optimality method in Reckase (2003; 2010). A comprehensive description of integer programming approach is provided in van der Linden (2005). Basically by this approach, severe constraints are imposed in the test content blueprint and other qualitative features of the items (e.g., item numbers, word count). However, these severe constraints may lead to statistically less optimal tests and pose security risks (Luecht, 2003). In addition, when the number of constraints is large, the test procedures are cumbersome, time-consuming or even infeasible (Zheng et al., 2012). Comparatively speaking, the *p*-optimality approach is easier to implement with success. By this method, items are sorted into a set of "bins" which are defined on the examinees' proficiency scale. The width of bins is determined by a series of factors, such as the desired information for the target test and the model used for item parameter calibration. The bins are designed to tally the number of administered items needed for the corresponding range in the proficiency scale. Currently many CAT programs, such as the National Council Licensure Examination and Armed Services Vocational Aptitude Battery are using the *p*-optimality method to design their item pools (He & Diao, 2014).

The integer programming approach in CAT was extended to an item pool design discussion for a three-stage MST design in Veldkamp (2014). However, although the

p-optimality method has been studied extensively in CAT (Reckase, 2003, 2010; Gu, 2007; Zhou, 2012; He & Reckase, 2014, Mao, 2014) for different item types using different IRT models, and in MST for a 1-2 panel design (Reckase, 2006) with a short test (e.g., 20- item test), it has never been investigated and applied to support various MST panel designs (e.g., 1-3, 1-2-2, 1-2-3) with different test specifications (e.g., different test lengths and routing test proportions). As van der Linden noted (2005), an item pool may serve an adaptive testing program best if the distribution of the maximum information across items in the pool follow the distribution of examinees' latent traits. An optimal item pool designed with the *p*-optimality method captures this essence. Thus one purpose of the study was to design optimal item pools to support different MST panel designs by extending the *p*-optimality method. Another purpose was to compare the different MST panel designs in terms of measurement accuracy for examinees' ability estimates. To develop an MST, test developers need to decide on many factors, such as total test length, routing test length, number of stages, number of modules within each stage, routing rules, module selection, ability estimation, and scoring. Based on the purposes mentioned above and the additional design decisions that must be made, this study compared the performance of MST assembled under the most popularly studied panel designs in the literature, including 1-2, 1-3, 1-2-2, and 1-2-3. It also used one of the most popularly investigated IRT models, the Rasch model, to generate simulated optimal item pools with and without practical constraint of exposure control. This study compared the performance of simulated and operational item pools in terms of measurement accuracy for the examinees' ability estimates.

The results from this research will provide information for efficiently designing optimal item pools for multistage computerized adaptive testing, facilitating the MST assembly process

and improving scoring accuracy. The study results will also inform operational item pools to design practical MST tests.

CHAPTER 2: Literature Review

This chapter first presents an overview of the theoretical background for test design in MST. It then provides background information for MST framework and components. The last section contains a general review of the literature for MST test designs, practical constraints in item selection, routing rules, test assembly, ability estimation and item pool design.

2.1 IRT models

Item response theory (IRT) is a model-based measurement approach developed for estimating examinees' abilities based on their responses to test items and on the properties of the items that are administered (IRT, Lord, 1980; de Ayala, 2009). IRT is classified into unidimensional and multidimensional models depending on the number of latent traits that are assumed. For convenience, this study was limited to unidimensional IRT models. Underlying unidimensional IRT is the assumption that there exists a location on an unobservable latent continuum for each examinee (usually denoted θ) which determines their probability of correctly responding to a given test item (Lord, 1980). Examinees' responses to one item in a test are assumed to be independent of their responses to other items after conditioning on this unobservable latent proficiency.

Commonly used IRT models include the one-parameter logistic model or Rasch model (1PL; Rasch, 1960); two-parameter logistic model (2PL; Birnbaum, 1968; Lord, 1980; de Ayala, 2009) and three-parameter logistic model (3PL; Birnbaum, 1968; Lord, 1980; de Ayala, 2009) based on the number of item parameters that are estimated for each item to model the relationship between examinees' probability of correctly responding to an item and their latent ability. Since Rasch models have several desirable measurement and psychometric properties, such as "observable sufficient statistics for the model parameters and a relatively small sample

size requirement for parameter estimation" (Wang & Wilson, 2005, p. 128), it was chosen to be used in the present study for item generation and calibration.

The Rasch model, as proposed by Georg Rasch (1960), is the simplest and one of the most commonly used IRT models for dichotomously scored items. It has one parameter to describe the examinees' ability and one parameter to describe the item difficulty. The equation for the model is:

$$P_{ij}(U_i = 1|\theta_j; b_i) = \frac{e^{1.7(\theta_j - b_i)}}{1+e^{1.7(\theta_j - b_i)}} \qquad (1)$$

where U_i is the item score of examinee *i* on item *j*, θ_j is the latent trait of person *j* and b_i is the item difficulty for item *i*. In the model, 1.7 is a scaling constant which places the parameters of the logit model onto the probit model scale (de Ayala, 2009). The equation denotes the probability of any examinee *j* responding correctly to any item *i*.

2.2 Item and test information

In IRT, item information is used to determine the measurement precision for different examinees' ability levels on the latent continuum. Birnbaum (1968) introduced Fisher's information (FI) to explain the information function for dichotomous items with the Rasch model:

$$I_i(\theta)=1.7^2\ P_i(\theta_j)Q_i(\theta_j), \qquad (2)$$

where $I_i(\theta)$ is the amount of information that item *i* provides for examinee *j*. $P_i(\theta_j)$ is the probability of person *j* correctly responding to item *i*, and $Q_i(\theta_j)$ is the probability of incorrect response. $I_i(\theta)$ has a maximum value of $1.7^2 * 0.25$ with the Rasch Model.

Since the test items are independent of each other, the information for the whole test is the summation of the individual item information. The information function for the whole test is:

$$I_L(\hat{\theta}_j) = \sum_{i=1}^{L} I_i(\hat{\theta}_j), \qquad (3)$$

where L is the test length and $I_L(\hat{\theta}_j)$ is the total test information at the estimated ability level of examinee j. An important feature from the definition of test information is that the more items included in the test, the greater the amount of test information. Thus, in general, tests with longer lengths will measure an examinee's ability with greater precision than will shorter tests. Another important feature of the test information is that the higher the test information is at a particular ability level, the more precise the test is in measuring the examinees at that ability level. This feature of test information can assist test developers to choose items that maximize information at particular cutoff points in the test for various test design purposes.

In the context of MST, the module selection from the item pool is decided on the maximum information it provides at the current ability estimate. As discussed in Weissman et al. (2007), the FI method for a dichotomous random variable x:

$$FI_j(\theta) = -\sum_{x\epsilon\{0,1\}} f_j(x|\theta) \frac{\partial^2}{\partial \theta^2} \ln f_j(x|\theta), \qquad (4)$$

where $f_j(x|\theta)$ is the probability of response x for item j conditional on θ. $FI_j(\hat{\theta})$ is evaluated prior to administering item j, and $\hat{\theta}$ is the provisional latent trait estimate. With the assumption of conditional independence of item responses given θ, the estimated FI for the module is the sum of the FI for all individual items in this module evaluated at the estimate $\hat{\theta}$. The equation is:

$$FI^b = \sum_{j\epsilon J_b} FI_j(\hat{\theta}), \qquad (5)$$

where b refers to the b^{th} module. At the examinees' estimated $\hat{\theta}$ (obtained after stage r has been completed), the routing occurs to direct examinees to the stage $r+1$ with the largest value of FI^b for the b^{th} module. Test information for any given examinee is the sum of in all modules the examinee experienced module information along his or her complete pathways.

Test information functions can also be used to evaluate the measurement precision at specific ability levels through calculating the conditional standard error of the measurement (SEM). The SEM for a given ability level (θ) is equal to the reciprocal of the test information function at the specific ability estimate, which is defined as:

$$SEM\ (\hat{\theta}_j) = \frac{1}{\sqrt{I_L(\hat{\theta}_j)}}, \qquad (6)$$

where $SEM\ (\hat{\theta}_j)$ refers to the standard error of measurement for the ability estimate of θ_j, and $I_L(\hat{\theta}_j)$ is the test information function at the estimated ability level of θ for examinee *j*. The overall SEM for the whole test is calculated through taking the average of all conditional SEMs across all θ points. The overall test information is calculated based on the overall SEM for the whole test.

For computer adaptive tests, the test information is calculated after the examinee's response to each item, so it can serve as a criterion for stopping the test once the desired conditional precision of measurement is achieved (Thissen, 2000). Because items vary across students in CAT, marginal reliability (Thissen & Wainer, 2001) was recommended to be reported instead of the internal consistency for the linear test. Marginal reliability is defined as:

$$r = (\sigma^2 - \sum_{i=1}^{N} SEM_i^{\ 2}/N)\ /\sigma^2, \qquad (7)$$

where *r* refers to the marginal reliability, σ^2 is the variance of ability estimates, *N* is the number of examinees. Marginal reliability is a measure of the overall reliability of the test based on the average conditional standard errors of the ability estimates for all the examinees.

2.3 MST components

The MST framework includes several basic components: modules, panels, stages, and pathways (Luecht & Nungester, 1998). An MST instrument generally begins with a first-stage module or routing test with a range of item difficulties covering the middle portion of the

distribution of examinee achievement. Examinees are adaptively routed to next stage(s) based on their performance in the previous stage.

2.3.1 Modules

Modules refer to sets of items which are administered and scored as a unit (Yan et al., 2014). Based on a test blueprint, modules usually follow specific content specifications and certain reliability and difficulty requirements. For example, the item difficulty for the same module is homogeneous and the range of difficulty for different modules is different. Variation of the item difficulty is allowed within modules, but the items for each module usually cover a portion of the entire range of examinee achievement. It is possible that the range of item difficulty across different modules at the same stage has an overlap since it will avoid problems of measurement accuracy with examinees whose ability levels are near the border between two modules. The size of modules may range from small to large, depending on the test specifications and requirements. But modules discussed here are not similar to "testlets" in Wainer and Kiely (1987) in which items are related to a single topic such as a reading passage. They are more generally referred as "bundles of items" given that the items in modules are assembled and administered together as a set (Ariel, Veldkamp, & Breithaupt, 2006).

2.3.2 Pathways

Based on the routing rules and examinees' performance, they are administered different modules at different stages. Pathways are the sequence of modules that individual examinees follow in the MST process. Each examinee only follows one pathway in the whole testing procedure.

2.3.3 Panels

Panels are the assembled modules that meet the test specifications in MST. A panel usually has several different pathways and multiple panels are needed to control for item and module exposure rate (Luecht & Nungester, 1998). To ensure comparability and reliability, the panels are intended to be parallel. That is, the structure of the panel (e.g., the number of stages, length of each stage, number of modules at each stage, and pathways from stage to stage) is identical across all parallel panels.

2.3.4 Stages

Stages are the administrative unit of the MST to allow for adaptation of the test to an examinee (Luecht & Nungester, 1998). The first stage often contains one module and later stages may contain two, three or more modules. Most MST designs have two, three or four stages (Yan et al., 2014). More stages allow for greater adaptation and better measurement results. However, efficiency of the test should also be considered since adding stages to the test will generally increase the complexity of the test assembly without necessarily adding much to the measurement precision (Luecht, Nungester, and Hadadi, 1996).

2.4 MST Research

2.4.1 MST test designs

Designing an MST involves as many components as creating a P&P test or an item-level CAT. Questions regarding test length, number of stages, number of modules within each stage, length of each module, number of paths between modules, routing rules, module selection, scoring, and the reliability and validity all need to be considered.

Using the three-parameter logistic (3PL) IRT model, Reckase (2006) explored a 1-2 MST design for a 20-item achievement test with exhaustive length combinations of the first-stage test

and the second-stage test. For example, the first-stage tests were simulated for lengths from 2 to 18 items, and the second-stage tests had complementary ranges of 18 to 2 items. The *p*-optimality method was applied to build up the optimal item pool to support the assembly of the 20-item test. Results showed that when the length of the second stage test was three times longer than the first stage one, the test worked the best. Since for the 20-item test, it was assumed that the item parameters were known without error and the 3PL model accurately reflects the interaction between the examinees and the items, to compensate for this strong assumption, the study was extended to a slightly longer test with 24 items. Combinations of test length were from 2 to 22 and the complementary range is from 22 to 2. The best combination was still achieved when the second-stage test was three times longer than the first-stage test.

Patsula & Hambleton (1999) studied a two-stage MST design (e.g., 1-3 and 1-5) and a three-stage MST design (e.g., 1-3-3 and 1-5-5) using the 3PL IRT model. The accuracy of ability estimates generated from MST were compared with that from P&P and CAT. Their study results showed that the number of modules at later stages did impact the measurement accuracy, and the designs with more modules at later stages had higher measurement accuracy than those with fewer modules.

Armstrong & Edmonds (2004) studied a variety of MST designs using the 3PL IRT model with both three-stage designs (e.g., 1-2-3) and four-stage designs (e.g., 1-2-3-4). Both types of the designs considered have 100% of the examinees taking the first stage modules and an even proportion of examinees taking the modules at later stages. For example, in the 1-2-3 design, 100% of the population took the module at Stage 1, and 50% of the population took the easy module and 50% of the population took the hard module at Stage 2, and 33% of the population took each of the three modules at Stage 3. The results showed that the three-stage

design had the most desirable measurement accuracy, and the four-stage design with four levels at the final stage surprisingly was less favorable than the three-stage design. The cost of using four-stage designs is the complexity in the routing algorithm and the item pool usage in terms of test assembly.

Zheng et al. (2012) compared the performances of two different MST designs (e.g., 1-2-3-4 and 1-2-4 designs) using automated top-down assembly strategy for a fixed-length large-scale classification test with a real item bank of 600 items. Both longer earlier stages (e.g., routing test) and longer later stages (e.g., the final stage in the MST) were included in the experimental conditions. Longer routing test length was designed to provide more precise routing of examinees to later stages, and longer later stages were intended to provide more items for accurate estimation when the test becomes more aligned with the examinee's ability level (Patsula, 1999; Zheng et al., 2012). However, similar to the previous research, the study results did not show the advantages of longer earlier stages over longer later stages or vice versa. Regarding the comparison between the two different MST designs, the results showed that the four-stage MSTs provided slightly higher correct classification rates than the three-stage tests. But no significant advantages for the four-stage tests were discovered.

Other MST designs that are were reported in the literature include a 1-2-2 design which discussed the impact of statistical constraints on classification accuracy in a licensure test (Park, 2013) and one which discussed examinee proficiency classifications in a language proficiency assessment test (Luecht, 2003), the 1-3-3 design (Hambleton & Xing, 2006; Jodoin et al, 2006; Davis & Dodd, 2003; Luecht, Brumfield, & Breithaupt, 2002, 2006; Zenisky, 2004), and the comparison of 1-3, 1-3-3, and 1-3-5 designs in the credentialing medical exam context (Luecht et al., 1998).

2.4.2 Practical constraints in item selection

In CAT, practical constraints, such as content balancing and item exposure control, need to be addressed in the item selection process for meeting test specifications and test security purposes. In the case of MST, however, test developers can preassemble modules to have better control over both statistical (e.g., item difficulty distribution) and nonstatistical attributes of items (e.g., content balancing) (Yan et al., 2014). Test developers may check on detail that "formal content specifications are met as well as that the informal nonexplicit content characteristics of items are appropriately represented and distributed" (Hendrickson, 2007, p47). This greatly reduces problems that may arise in complex item selection algorithms. For example, in the Uniform CPA exam, multiple panels are simultaneously constructed prior to test administration so that the equivalence in psychometric properties is ensured and overexposure of highly discriminating items is reduced across panels (Melican, Breithaupt, & Zhang, 2010). Content balancing is achieved by assigning particular portions of a total test blueprint to specific modules so that the contents required in a test blueprint are representative when the examinees complete the final stage of the test (Zhang et al., 2006).

Exposure control in MST, as discussed in Yan et al. (2014), may incorporate the features of linear tests or CAT. For example, as in linear tests, it is possible that the MST panels retire after being administered over a certain period of time, and then be returned to the item pool and reassembled for future usage. The CAT approach may require constructing multiple item pools for parallel modules at different stages of MST, and at each stage, the modules may be randomly drawn from the module pool. Compared with CAT, the exposure control for MST is much simpler because only random selection of modules is required to achieve both conditional and unconditional exposure control at preset levels (Yan et al., 2014).

In Zheng et al.'s study on MST (2012), to make the exposure rates for different modules more uniform and more efficiently use the item bank, the number of forms to be assembled for each module was made inversely proportional to the number of modules of the stage that it belonged to. The assumption is that there is a roughly equal proportion of examinees being routed to different modules. Edwards et al. (2012) introduced a uniform exposure control for multistage computerized adaptive test. By this method, the number of routing blocks is made parallel with the levels in the subsequent stages, and examinees are randomly assigned to those routing blocks. The uniform item exposure can then be accomplished by appropriate selection of cutoff scores and routing of examinees to subsequent stages.

Luecht (2003) proposed a different method with the bundled multistage adaptive testing (BMAT) framework which permits varying levels of exposure control at different stages and random assignment of module for each examinee. In the BMAT framework, item exposure mechanisms are built into the module pre-construction process. By this method, there are "primary" and "auxiliary" routes through the blocks used by the majority and minority of examinees.

Under IRT, one statistical constraint is target test information (TIF). The TIF determines the amount of information at either module level or pathway level along the examinees' proficiency scale. Higher TIF corresponds to higher level of measurement precision since it is inversely related to the standard error of measurement (Emberston & Reise, 2000). Researchers who studied the relationship between TIFs and classification accuracy showed that accuracy increased with increase in TIFs, especially for the module at the first stage (Zenisky, 2004; Kim et al. 2012; Park, 2013). Target TIFs are difficult to assess before the test construction.

However, the pathway information could be predicted depending on the average item information in the pool and the number of panels under construction (Park, 2013).

2.4.3 Routing rules

Three routing methods were documented in Weissman, Belov & Armstrong (2007): number correct (NC), maximum Fisher information (FI) and maximum mutual information (MI). For the NC method, a lookup table containing threshold scores is predetermined before an MST is administered. The module an examinee is routed to is dependent on whether the provisional estimate is smaller or larger than the threshold score (Weissman, 2014). For the FI method, module information is calculated at the provisional estimate of the examinee's proficiency, and the module with maximum information at the provisional proficiency score is assigned to the examinee. Using the maximum mutual information method (MI), the maximum information is calculated after the examinees complete stage *r* using their posterior ability distribution with a uniform prior, probability of response, and the marginal response distribution (detailed equations see Weissman et al., p. 10). The MI method provides "a measure of the expected reduction in uncertainty in predicting a person's proficiency level" (Weissman et al., p. 8).

Research results indicated that NC routing utilized the MST modules and pathways most uniformly with the observed routing percentages for modules matching closely to those expected (Weissman et al., 2007). However, FI and MI routing methods led to higher correct classification rates of larger percentage of test takers (88.8% for FI and 89.1% for MI) overall than NC routing (85.3%). Although the MI routing method is similar to the FI method in classification of test takers, it only unutilized one out of the four paths for all test takers in Weissman et al.'s study (2007). The paths utilized by the FI method and NC method were not significantly different from each other except for path one.

As discussed in Weissman (2014), if the focus is on the individual examinee, routing rules based on information function should be utilized. If the focus is on the group, the routing rules should take into account the distribution of proficiencies in the population and information function.

2.4.4 Test assembly

Multiple panels are required to be developed in the process of test assembly in order to ensure test security. Generally speaking, assembling MSTs can be conducted either through self-directed programming (Davis & Dodd, 2003; Keng, 2008; Kim et al., 2008) or by using computer software such as an MST automated test assembly (ATA) program (CASTISEL, Luecht, 1998; Zenisky, 2004). ATA computer software applies optimization algorithms with either linear programming or heuristics to simultaneously construct multiple panels. Linear programming is based on a mathematical modeling of assembly which requires strict test assembly constraints (e.g., item exposure, target TIFs, and content coverage) (Adema, 1990; Armstrong et al., 2004; Luecht et al., 2006; Armstrong & Roussos, 2005; Luecht & Nungester, 1998). Comparatively speaking, the heuristic approach does not guarantee that all constraints are satisfied, but it is less computationally intense and also widely used in research studies (Luecht, 1998; Patsula, 1999; Hambleton & Xing, 2006; Jodoin et al., 2006; Zheng et al., 2012; Kim et al., 2012).

According to Luecht and Nungester (2000), MST panels can be assembled in three ways: bottom-up assembly, top-down assembly, and mixture assembly. In the bottom-up assembly, parallel test forms for modules at certain stages are first assembled. The module-level specifications, such as the statistical characteristics, test information, and content constraints, are all addressed so that modules can be mixed and matched in each stage to form panels and used

interchangeably across panels. In the top-down strategy, one optimization procedure is needed to ensure test-level specifications, such as the target TIFs specified for each of the pathways. In this method, the assembled test forms for modules at each stage are not parallel, so they must be combined in prescribed pathways to fulfill the target test specifications. Modules assembled using the top-down strategy at certain stages cannot be used interchangeably across panels. The mixture assembly specifies both the module-level and test-level constraints. For example, it satisfies some of the test specifications at the module level and others at the test level.

2.4.5 Ability estimation

The ability estimation procedures used in CAT are also applicable in MST. Typically two methods can be used to estimate the person parameters: Maximum likelihood estimation (MLE) (Lord, 1980; Birnbaum, 1968) and a Bayesian method (Owen, 1975). In MLE, the examinee's most likely position on the latent trait is located through maximum likelihood estimation. In other words, MLE is a procedure of finding the value of desired parameters that make the observed data distribution the most probable (de Ayala, 2009). This method is presented below:

$$L\,(\boldsymbol{x_j}|\theta_j;\boldsymbol{b}) = \prod_{i=1}^{L} P_i^{\,x_{ij}}(1-P_i)^{(1-x_{ij})}, \qquad (8)$$

where P_i is short for $P_{ij}\left(U_i = 1\middle|\theta_j; b_i\right)$, x_{ij} is person j's response to item i. $\boldsymbol{b}$ is a vector containing item location parameters, L is the number of items on the test. Since with the increase of the number of items, the product of the probabilities becomes too small, the natural logarithmic transformation of the probability is typically used. The utilization of logs results in a likelihood that is called log likelihood function, $\ln(\boldsymbol{x_j})$, where

$$\ln L(\boldsymbol{x_j}|\theta_j;\boldsymbol{b}) = \sum_{i=1}^{L}(x_{ij}\ln(P_i) + (1-x_{ij})\ln\,(1-P_i)) \qquad (9)$$

When both person and item parameters are unknown, joint maximum likelihood estimation method are used to simultaneously determine these parameters that maximize the joint likelihood

of the observed data (de Ayala, 2009). The marginal maximum likelihood estimation method is another special case of maximum likelihood estimation (Bock & Lieberman, 1970; Bock & Aitkin, 1981), in which one parameter is estimated by maximizing the marginal likelihood function through integrating out another parameter.

With Bayesian methods, a prior distribution of the examinees, for example, a normal distribution, is usually assumed for the ability parameters. A posterior distribution of the ability parameters is obtained with the help of the likelihood of the observed data given the specified IRT model. When the mode of the posterior distribution is used as the final ability estimate, it is named maximum a posterior (MAP); when the mean of the posterior distribution is used as the final ability estimate, it is called an expected a posteriori (EAP) estimate (de Ayala, 2009; Embretson & Reise, 2000).

2.4.6 Item pool design

An item pool is defined as consisting of "a maximal number of combinations of items that (a) meet all content specifications for the test and (b) are most informative at a series of ability levels reflecting the shape of the distribution of the ability estimates for a population of examinees (van der Linden, Ariel, and Veldkamp, 2006, p82). A successful MST assembly is inseparable from an optimal item pool that provides sufficient items (Luecht & Nungester, 1998).

Xing & Hambleton (2004) studied the impact of item pool quality on the accuracy of ability estimation and concluded that the better (e.g., higher item discrimination) and bigger the item pool, the more information the MST design provided across the wide range of ability. The results from Wang, Fluegge, & Luecht (2012) also showed that the quality of item pool was the primary factor that impacted the efficiency of MST design. An item pool that was specifically

designed for a MST dramatically improved scoring accuracy. Item pool design is meant to develop a blueprint in which the statistical and non-statistical attributes of items are described (Veldkamp, 2014). Statistical attributes include the item selection algorithm, exposure control procedure, termination procedure, and item overlap restriction. The non-statistical attributes deal with such issues as target examinee population distribution characteristics, content balancing and test purpose (He & Reckase, 2014; Veldkamp, 2014). For CAT, two approaches have emerged for item pool design: the integer programming approach, represented by the shadow-test approach in Veldkamp & van der Linden (2000), and the heuristic approach, as represented by the *p*-optimality method in Reckase (2003; 2010).

2.4.6.1 Integer programming approach

The initial study of item pool design using integer programming was addressed by Boekkooi-Timminga (1991) in which integer programming was used to calculate the number of items needed for future test forms. A sequential approach was used to maximize the test information function under the Rasch model and the results were then applied to improve the composition of the item banks.

To reduce item exposure and enhance item use efficiency, Stocking and Swanson (1998) applied an optimal design method for assigning items from a master bank to a set of generated banks, including independent banks and banks with overlapping items, using the three-parameter logistic (3PL) model and weighted deviation method in selection of items. The test assembly to desired content and measurement properties was achieved using a standard linear programming model. More detailed explanations of this model are described in Stocking and Swanson (1993).

The definitive work for integer programming in optimal item pool design is found in van der Linden, Veldkamp and Reese (2000). The method was intended to create a blueprint for an

item pool which stipulates the attributes of the items, including both categorical constraints (e.g., item content, cognitive level, format, author, answer key) and quantitative constraints (e.g., word counts, exposure rates, TIFs, expected response times, and item difficulty and item discrimination) required for the assembly of a pre-specified number of test forms from the pool. The blueprint is optimal in the sense that the effort or "cost" of item pool creation is minimized (van der Linden et al., 2000), and the number of unused items in the pool is minimized. Using the 3PL model, the method was demonstrated through designing a new item pool for the Law School Admission Test (LSAT) which consists of both item sets with a common stimulus and discrete items. With the cost function being built from a previous item pool of 5,316 items, the new optimal item pool was designed to support 30 test forms with no overlapping items across different forms. The strategy used to solve the item assignment models was through the simplex algorithm, which was implemented in the Consolve module in the test assembly software package ConTEST (Timminga, van der Linden & Schweizer, 1996). It was suggested that modifications are needed for the application of integer programming models to the actual testing programs when the test specifications are varied.

Extension of the integer programming approach to MST item pool design was discussed in Veldkamp (2014). Under the context of MST, several objective functions in the integer programming models were to be optimized for designing the optimal item pool, assuming the number of modules, and all the categorical and quantitative constraints are known. Objective functions, such as minimization of the costs of item writing, minimization of the number of items, and minimization of item overlap between modules, were all considered. New logical constraints, such as the relationship between items (e.g., item enemies), item sets (e.g., items belonging to a common reading passage), and item overlap among modules, needed to be added

while presenting the blueprint design for the optimal item pool in MST. It should be noted that the blueprint of the item pool does not need to be a static entity. In the process of item pool maintenance, the newly added items might deviate from the blueprint because only part of the attributes can be controlled by the item writers. To correct for these deviations, the item pool blueprint needs to be updated on a regular basis. The author introduced an index that could be used to denote the iteration of the blueprint in the MST context (more details see Veldkamp, 2014).

2.4.6.2 *p*-optimality approach

The *p*-optimality approach for designing optimal item pools in CAT was demonstrated using the Rasch model by Reckase (2003, 2010). In Reckase (2003), the study was limited to a fixed-length CAT program with no exposure control and content balancing. In Reckase (2010), the study was extended to a variable-length CAT program considering both exposure control procedures and content balancing. As defined in Reckase (2010), by optimal item pool, it refers to the fact that "whenever the CAT item selection algorithm is searching for a test item to administer, exactly the item that is desired is available in the item pool" (p. 129). With the Rasch model, an optimal item pool is the one that has a *b*-parameter exactly equal to the current θ estimate for every item selection. Routinely the size of the item pool is as large as 2^n-1, where *n* is the number of items administered to an examinee. For a CAT test with 20 items, the optimal item pool would need to contain 1,048,575 items, which is not practical for any item pool design. To make the concept of an optimal item pool realistic, Reckase (2003, 2010) applied the *p*-optimality approach or *r*-optimality approach (He & Reckase, 2014). By this method, the difference in the amount of information provided by an item about the current θ estimate when there is an exact match with the *b*-parameter and when there is not an exact match is considered.

Reckase (2010) illustrated the information function for a test item as fit by Rasch model for the situation of 90% of the maximum possible information. Suppose the item that is available for selection has information that is within 95% of the maximum possible, the selected item is within about .35 of exactly matching the θ value. This specifies a range and the criterion was called range optimality. The goal is to be *r*-optimal (e.g. .35-optimal). If an item pool meets the criterion of always having items available for selection that are 95% or more of the maximum possible information, the item pool is called .95 *p*-optimal. The specification of *p* in *p*-optimality can be used to determine the value of *r* for *r*-optimality (Reckase, 2010). The bolded horizontal line in the following figure refers to the level of 95% maximum possible information. The dotted lines refer to the unit which contains the range of .35 on each side of the mid-point of 0. If the item difficulty is .35 logit away from the current ability estimate, the most loss of information for the examinee with the selected item is 5% compared with a perfect match between the selected item and the current ability estimate.

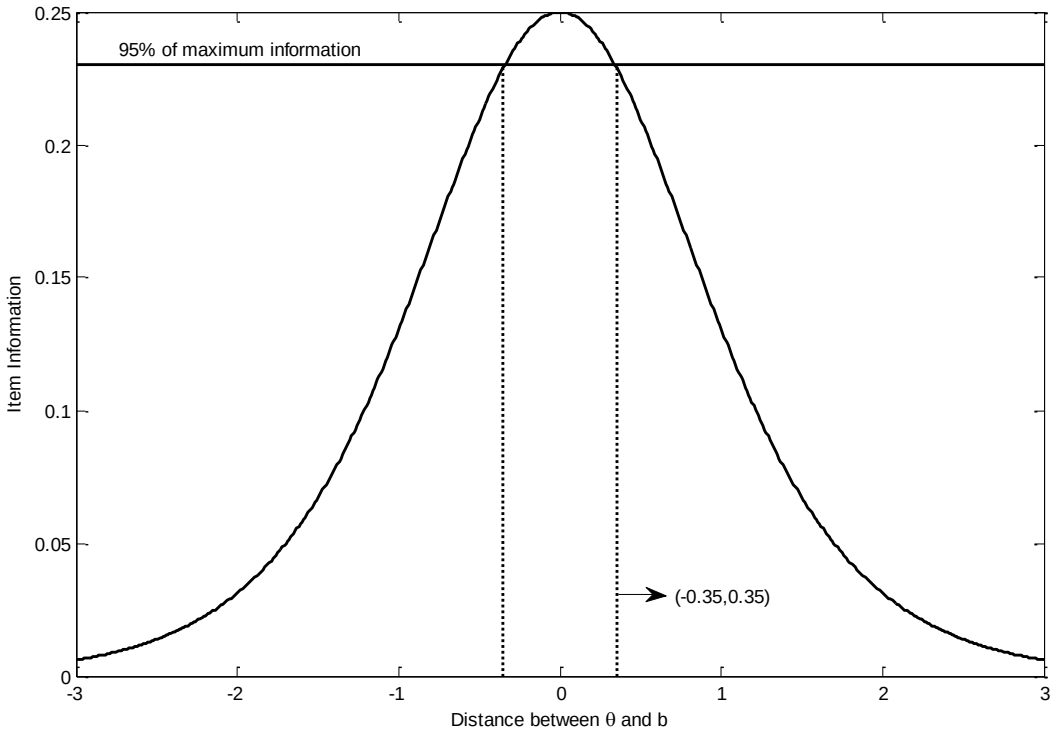

Figure 2.1 Item information function specified by a Rasch model

If *p* is smaller than 0.95 (farther away from 1.0), the width of the bin will be wider. For a .85 *r*- optimal item pool, the item selected is within about .85 θ-units of the desired item, or the full range of acceptable items is 1.7. In this study, the θ-units of the desired item, for example, .35, is termed a bin. To be consistent with the previous research, the term *p*-optimality was used to imply the optimal item pool.

The basic idea of implementing the *p*-optimality approach for optimal item pool design is to randomly select examinees from a target population. The selected items are sorted into a set of "bins" which are defined on examinees' proficiency scale, for example, bins of width .35 as discussed above. The width of the bins is determined by the desired information for the target test (e.g., 90% maximum information for the selected test item or 95% maximum) and the model used for item parameter calibration (e.g., 1PL, 2PL or 3PL models) (Reckase, 2003; 2010). After

one examinee takes the simulated CAT test, the optimal item sets for him/her are allocated into bins according the optimality criteria. Another examinee is then selected and the same procedure is repeated. The minimum common item sets for these individual examinees are determined by taking the union of the two individual item sets. The process continues until the number of the items in the union reaches an asymptote (Reckase, 2010), which also constitutes the desired optimal item pool size. By this method, the bins are designed to tally the number of administered items needed for the correspondent range in the proficiency scale. Examinees with similar proficiency levels may share similar sets of items within the same bins. Thus the desired number of items in the pool is reduced.

To show the practicality of the *p*-optimality approach, Reckase (2010) applied it to design the optimal item pool for a variable-length CAT test in certification/licensure using the Rasch model. With eight content areas involved in the test, the items were tallied into bins for each content area and 15 items were contained in the optimal item set at each bin to account for the exposure control requirement. The total size of the items for all content areas is 1,602, which is smaller than the operational item pool size of 2000. To check the performance of the simulated item pools, two *r*-optimal pools with variable bin widths were compared with the operational item pool and conditional standard errors were computed at equally spaced points along the θ scale. The results showed that in the middle near the cutoff score, the different item pools resulted in equal precision of estimates. However, the conditional standard errors increased for the operational item pool at the two extremes while those for the simulated optimal item pools remained constant. The operational item pool had more items in the middle than necessary and less spread of item difficulty than the *r*-optimal pools.

He and Reckase (2014) provided additional details of applying the *p*-optimality method to design optimal item pools for a variable-length licensure CAT program using the Rasch model. Seven candidate item pools were designed to demonstrate how design factors, such as variable bin widths, exposure control, and content balancing, affected item pool characteristics and the performance of the CAT program. The results indicated that the item distributions across all the content areas were similar. More specifically, the item distributions were negatively skewed with the peak centering on the middle category of *b* values near the cutoff score; the number of items in the bins father away from the cutoff score decreased; all the item pools covered a wide range of item difficulty levels along the ability scale and provided sufficient items to support the exposure control procedures. In addition, the item exposure rate distribution for each content strand also shared a similar pattern. Comparatively speaking, the item pools with exposure control were found to yield slightly more accurate classification accuracy and higher rate of underexposed items than those with no exposure control, and the their item pool size increased by about threefold. The item characteristics of the pool were affected by bin width, for example, the narrower the bin width, the larger the item pool size. To evaluate the feasibility of the simulated item pools, the performances of two selected pools were compared with that of a retired operational item pool in the same CAT program. The comparison results showed that the operational item pool yielded slightly longer tests on average, slightly lower classification accuracy, and higher biases and MSEs at the extreme abilities. But the operational item pool demonstrated a better item usage than the simulated pools with a lower percentage of underexposed items although at the cost of 400 more items.

The *p*-optimality approach was extended to the use of 3PL model in Gu (2007) for a CAT program (e.g., Armed Services Vocational Aptitude Battery, ASVAB) composed of dichotomous

items. Practical constraints, such as Sympson-Hetter exposure control method, were applied in the study for test security purpose, but content balancing was not considered. The results suggested that the simulated optimal item pools performed better than the operational pools no matter whether the exposure control procedure was implemented or not. More specifically, item pools designed with the *p*-optimality method had more items evenly distributed over a wider range of ability levels, which resulted in a better estimation accuracy at most latent ability levels. In addition, the simulated item pools had significantly smaller percentages of under-exposed items than the operational item pools.

Zhou (2012) explored the impact of the practical constraints, such as *a*-stratified exposure control method (with a maximum rate of .20) and content balancing, on the optimal item pool design for a polytomous CAT using generalized partial credit model (GPCM). The results indicated that the practical constraints of the exposure control and content balancing had little effect on the item pool size. However, the *a*-stratified exposure control affected the item pool characteristics. For example, the items in the simulated item pools without the constraint had larger *a*-parameters than those with the constraint. This resulted in the larger average maximum information as provided by those items and much larger item pool information than those with the constraint. Contrary to exposure control, the content balancing had little impact on the item pool design for the polytomous items. In the conditions where only the content balancing constraint differed, the distributions of the *a*- and *b*-parameters were quite similar. The evaluation results revealed that the optimal item pools designed with the *p*-optimality method supported the polytomous CAT implementations with regards to the measurement accuracy and item pool usage. More specifically, enough items were provided in the simulated item pools

which had maximum information across the entire ability continuum and the *a*-parameters spanned evenly under the *a*-stratified method constraint condition.

Mao (2014) applied the *p*-optimality method in designing optimal item pools for a multidimensional CAT program using the multidimensional Rasch model. A total of 24 *p*-optimal item pools were designed and then developed based on different test specifications, with different correlations among dimensions, different bin sizes and under different exposure control conditions. Content balancing was not considered. Since an operational multidimensional item pool in CAT does not exist in practice, the baseline item pools were created following those used in research articles. The results indicated that the simulated optimal item pools performed similarly with the baseline item pools in terms of the measurement accuracy for ability, but they contained fewer items (e.g., more than 100 items smaller) and had better item pool usage. Similar to Gu (2007) and Zhou (2012), the size of the simulated optimal item pools for MCAT was sufficient for a large number of examinees and the items in the pools spanned the entire range of item difficulty, which resulted in a good estimation accuracy at most ability levels. Item pool size was found to be related with several factors: the bin sizes, the test specifications, correlations among dimensions, and the exposure control condition. For example, under the exposure control condition, a larger item pool was necessary when the bin size increased, the test became non-simple structure, and the dimensions were highly correlated.

2.5 Statement of the problem

The review of recent literature on MST design reveals that popularly investigated designs include two-stage, three-stage, and four-stage designs. It is concluded from the literature that although the four-stage designs with four levels at the final stage provided slightly higher correct classification rates than the three-stage designs, they only brought negligible increase in

estimation and scoring accuracy (Zheng et al., 2012; Jodoin et al., 2006-zheng et al). In some studies (e.g., Armstrong & Edmonds, 2004) the four-stage designs were even less favorable than the three-stage designs. Comparatively speaking, the cost of using four-stage designs was to bring in greater complexity in the routing algorithm and poorer item pool usage (Luecht & Nungester, 1998; Patsula & Hambleton, 1999; Zenisky, Hambleton, & Luecht, 2010).

In terms of MST panel design, Reckase (2006) addressed the ideal test configurations for a 20-item or 24-item test especially regarding the proportion of the first-stage and second stage test length. But the study was only limited to short test length for a particular 1-2 design, the ideal MST configurations for other test lengths (e.g., medium and long), and for other panel designs (e.g., 1-3, 1-2-2, and 1-2-3) have not been explored and identified. Some studies (Zheng et al., 2012; Chang & Ying, 1996) investigated whether longer earlier stages and longer later stages in MSTs would bring differences in the measurement accuracy of the ability estimation, but no studies were conducted to address the ideal proportions of the test lengths at the module level for various MST panel designs.

As shown in Armstrong and Edmonds (2004), the efficiency of any MST design relies on high item pool quality and effective utilization of its pool. A valuable resource for any testing agency is an ideal test blueprint which helps to design and maintain their item pool. As evidenced from the literature review that the implementation of the integer programming approach in item pool design relies heavily upon specialized knowledge of linear programming and specific software, such as CPLEX and LINDO (He & Reckase, 2014). More importantly, severe constraints are imposed in the test content specifications and other qualitative features of the items (e.g., item numbers, word count). One consequence of these severe constraints is that they may lead to statistically less optimal tests and pose security risks (Luecht, 2003). In

addition, when the number of constraints is large, the test procedures are cumbersome, time-consuming or even infeasible (Zheng et al., 2012). Comparatively speaking, the *p*-optimality method is easier to implement and its application does not require any knowledge of a specialized sub-field of psychometrics and software. Currently many CAT programs, such as the National Council Licensure Examination and Armed Services Vocational Aptitude Battery use the *p*-optimality method in designing their item pools (He & Diao, 2014).

Although Reckase (2006) applied the *p*-optimality method to support a MST 1-2 panel design for a short test using 3PL model, we are not clear whether this method is also feasible in supporting other MST panel designs (e.g., 1-3, 1-2-2 and 1-2-3) containing medium and long tests and different routing test proportions. The feasibility of applying the Rasch model in optimal item pool design using this method is also unknown. As more large-scale educational assessments are moving towards the MST design, it will be helpful to investigate the measurement accuracy provided by the various MST designs and inform possible optimal test design choice.

Using one of the most popularly investigated IRT models, the Rasch model, one purpose of the present study was to enhance item pool utilization by the *p*-optimality method to support different MST panel designs, including 1-2, 1-3, 1-2-2, and 1-2-3. Item pools designed with and without exposure control were explored and item parameter characteristics were compared. Another purpose of the study was to evaluate the performances of various MST designs with different test configurations and see how the test lengths and routing test lengths impacted the measurement accuracy. Since the study was primarily interested in a test designed to have a good estimation across a range of ability levels, the target test proposed is an achievement test. For generalizability purpose, different test lengths (e.g., short, medium and long test) were all

included in the study. More specifically, different test lengths (e.g., 20-, 40-, 60-item test) with different proportions of the first stage (20%, 30% and 40% of the test) and the correspondent second-stage and third-stage tests were all studied. The performances of the simulated item pools were evaluated against many evaluation criteria in terms of measurement accuracy (e.g., bias, Root Mean Squared Error (RMSE), classification accuracy, item exposure rate, and the correspondent conditional evaluation statistics etc.). The *p*-optimality method was used to design an optimal item pool in an operational MST context to see if it was feasible to support the operational test. The following research questions were addressed in the present study:

1) Will the *p*-optimality method achieve sufficient measurement accuracy when used to design optimal item pools to support the different MST panel designs under different test specifications (e.g., different test lengths and routing test proportions), and is this method applicable to operational item pool design?
2) Will the test length and routing test proportions (e.g., the proportion of the routing test length to overall test length) have any impact on the measurement accuracy for all MST designs?
3) How does the practical constraint of exposure control affect the features (e.g., item pool size and parameter characteristics) of the optimal item pools to support the different MST panel designs?

CHAPTER 3: Methodology

This chapter first introduces the optimal item pool design for the different MST panel designs in the study. Then it discusses the test development, including test designs, routing methods, practical constraints in item selection, such as exposure control, and test assembly methods. In the research design section, a simulation study design and procedure, together with the evaluation criteria for item pool performance are described. This section also introduces how the operational item pool is formed and compared with the performance of the simulated optimal item pools.

3.1 MST test development

3.1.1 MST designs and test configurations

For generalizability purpose, the simulated test lengths include short, medium and long tests with 20, 40, and 60 items. The MST designs explored are among the most popularly investigated ones in the literature, which include both two-stage and three-stage structures: MST 1-2 design, MST 1-3 design, MST 1-2-2 design, and MST 1-2-3 design. In the MST 1-2 design, 1 means one routing module at stage one with medium level items, 2 means two modules at stage two (one module with easy items and one module with difficult items). In the MST 1-3 design, 1 means one routing module at stage one with medium level items, and 3 means three modules at stage two (one module with easy items, one module with medium level items, and one module with difficult items). In the MST 1-2-2 design, 1 means one routing module at stage one with medium level items, the first 2 means two modules at stage two (one module with easy items, and one module with difficult items), the second 2 means two modules at stage three (one module with easy items, and one module with difficult items). In the MST 1-2-3 design, 1 means one routing module at stage one with medium level items, 2 means two modules at stage two

(one module with easy items, and one module with difficult items), and 3 means three modules at stage three (one module with easy items, one module with medium level items, and one module with difficult items). Since previous research suggested that four-level modules and more than three-stage panels provided a negligible increase in scoring accuracy (Armstrong & Edmonds, 2004), and only increased the complexity of test assembly (Luecht & Nungester, 1998; Patsula & Hambleton, 1999; Zenisky, Hambleton, & Luecht, 2010), panels with more than three modules at each stage and more than three stages were not included in the study. The detailed design structures are shown in the following figures:

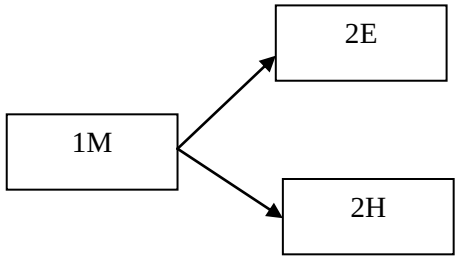

Figure 3.1 MST 1-2 design

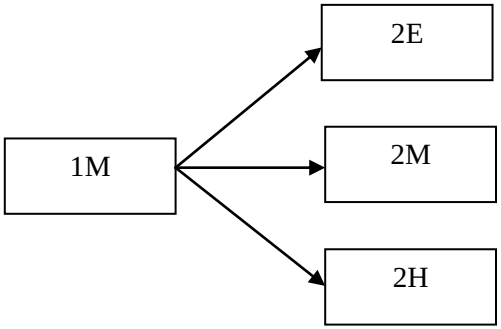

Figure 3.2 MST 1-3 design

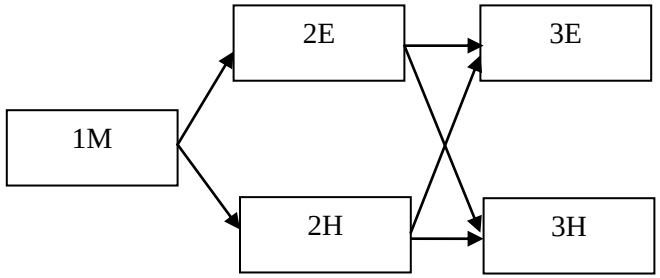

Figure 3.3 MST 1-2-2 design

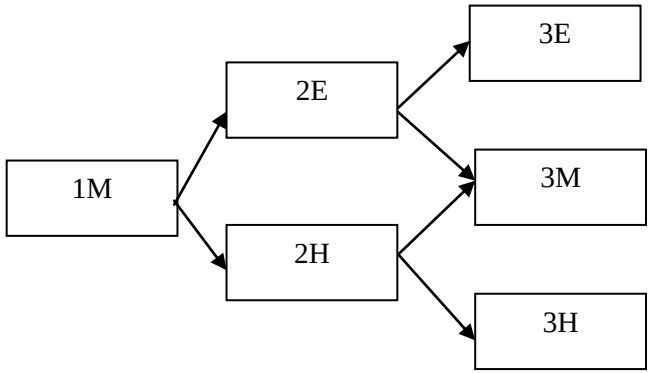

Figure 3.4 MST 1-2-3 design

In these figures, 1 means stage one, 2 means stage two, and 3 means stage three. M means items with medium item difficulty, E means easy items, and H means hard items. Given the examinees' abilities are unknown at the beginning of a test, the length of the routing test is particularly significant. Various proportions of the routing test length are considered in the study to see how this variation impacts the ability estimation accuracy. For the MST 1-2 design and MST 1-3 design, the proportions of the routing test length to the total test lengths are: 20%, 30% and 40%. The proportions of the second stage modules are: 80%, 70% and 60% of the total test lengths respectively. For MST 1-2-2 and MST 1-2-3 designs, the proportions of the routing test length to the total test lengths are: 20%, 30% and 40%. The correspondent proportions in the second stage modules are 40%, 30% and 20%, and the proportions for the modules at the final stage are: 40%, 40% and 40%. The rationale for splitting the proportions of the modules at the second and third stage is to ensure that enough items are allocated at the final stage for measurement accuracy purpose. For designs with no exposure control, the allocation of items for all the designs is described in Table 3.1 below. Take the MST 1-2-2 design as an example, for the test length of 40 (e.g., 1-2-2_n40) and routing test proportion of 40%, 16 items are allocated for stage one module, 8 items are allocated for each of the two modules at stage two respectively, and 16 items are allocated for each of the two modules at stage three. The allocation of items under exposure control conditions are described in Table 3.2 in 3.1.3.

Table 3.1 MST designs and the number of items across different stages

Designs	Proportions	Stages		
		1	2	3
1-2_n20	20%	4	16 (16)	
	30%	6	14 (14)	
	40%	8	12 (12)	
1-2_n40	20%	8	32 (32)	
	30%	12	28 (28)	
	40%	16	24 (24)	
1-2_n60	20%	12	48 (48)	
	30%	18	42 (42)	
	40%	24	36 (36)	
1-3_n20	20%	4	16 (16, 16)	
	30%	6	14 (14, 14)	
	40%	8	12 (12, 12)	
1-3_n40	20%	8	32 (32, 32)	
	30%	12	28 (28, 28)	
	40%	16	24 (24, 24)	
1-3_n60	20%	12	48 (48, 48)	
	30%	18	42 (42, 42)	
	40%	24	36 (36, 36)	
1-2-2_n20	20%	4	8 (8)	8 (8)
	30%	6	6 (6)	8 (8)
	40%	8	4 (4)	8 (8)
1-2-2_n40	20%	8	16 (16)	16 (16)
	30%	12	12 (12)	16 (16)
	40%	16	8 (8)	16 (16)
1-2-2_n60	20%	12	24 (24)	24 (24)
	30%	18	18 (18)	24 (24)
	40%	24	12 (12)	24 (24)
1-2-3_n20	20%	4	8 (8)	8 (8,8)
	30%	6	6 (6)	8 (8,8)
	40%	8	4 (4)	8 (8,8)
1-2-3_n40	20%	8	16 (16)	16 (16,16)
	30%	12	12 (12)	16 (16,16)
	40%	16	8 (8)	16 (16,16)
1-2-3_n60	20%	12	24 (24)	24 (24,24)
	30%	18	18 (18)	24 (24,24)
	40%	24	12 (12)	24 (24,24)

Note: 1-2_n20 means MST 1-2 design with a test length of 20 and the rest of the designs has similar representations; the number in parenthesis means the number of items at the same stage for another/other module(s).

3.1.2 Routing

Given the higher correct classification rate by the maximum Fisher information method (FI) discussed in Weissman (2014) and its popularity in MST research (e.g., Zheng et al., 2012; Luecht, Brumfield, & Breithaupt, 2002), the FI method was used as the routing method in the present study. As discussed in Weissman (2014), if the focus is on individual examinee, routing rules based on information functions should be utilized. If the focus is on the group, the routing rules should take into account the distribution of proficiencies in the population and information functions. The routing rules that are used in the present study adhere to the following: the simulated routing test is targeted at providing maximum information at the cutoff score determined by the distribution of examinees in the target test so that equal proportion of examinees could be routed to the next stage. In the simulation study, the distribution of the target examinees is assumed normal with mean of 0 and variance of 1. For the interim-stage modules in MST, items are selected to provide maximum information at the true ability of examinees and an equal proportion of examinees are routed to the next stage.

More specifically, with the Rasch model, the maximum information for the individual item is achieved when its item difficulty parameter matches the examinee's latent ability. For the MST design following a 1-2 structure, one cutoff score is used to route half of the examinees to the easy module and half to the difficult module at the second stage. The cutoff score is determined based on the mean of the distribution of simulated examinees in the study. For the 1-3 design, two cutoff scores are used to route one third of the examinees to one of the modules at the second stage. The two cutoff scores are determined based on 33^{th} percentile and 66^{th} percentile of the distribution of simulated examinees so that an equal proportion of examinees are routed to the second stage modules. Three-stage MST designs are more complicated. For

example, for the 1-2-2 design, the routing from the first stage to the second stage is the same with that of the 1-2 design, but two cutoff scores are generated at the second stage for routing an equal proportion of examinees to one of the two modules at the final stage. The two cutoff scores are determined based on the observed score variance instead of the true score variance of the simulated examinees. More specifically, the two cutoff scores are determined based on the 25^{th} percentile and 75^{th} percentile of the observed score distribution of simulated examinees at the second stage, which results in an equal proportion of examinees being routed to each of the modules at the final stage. Similarly, for the 1-2-3 design, two cutoff scores are generated based on the 33^{th} percentile and 67^{th} percentile of the observed score distribution of simulated examinees at the second stage, and an equal proportion of examinees are routed to one of the three modules at the final stage.

3.1.3 Exposure control

To maintain test security and module exposure control, the procedure for developing multiple panels and parallel modules in the study followed the method discussed in Zheng et al. (2012). The number of forms to be assembled for each module is inversely proportional to the number of modules of the stage it belongs to when an equal proportion of examinees are routed to different modules at later stages. By test forms in the study is meant the parallel modules at each stage. The numbers of items for each module under exposure control conditions are multiples of the ones as shown in Table 1 above. The number of test forms used for the exposure control conditions are shown below.

Table 3.2 The number of test forms in different MST designs with exposure control

	Stages		
Designs	1	2	3
1-2	18	9 (9)	
1-3	18	6 (6) (6)	
1-2-2	18	9 (9)	9 (9)
1-2-3	18	9 (9)	6 (6) (6)

*Note: T*he number in parenthesis means the number of test forms at the same stage for another/other module(s).

The number of test forms is determined by considering several factors. First, the number of test forms must be whole numbers for each module at each stage; second, since there is one module at the first stage for every MST design, the number of test forms for the initial stage is set as the same; third, since the inverse proportion method is applied in the study for exposure control and fewer test forms would occur at the interim as well as the final stages, the minimal number of test forms required for the second stage as well as the final stage for the various MST designs are determined using pilot studies. The factors, such as the exposure rate, the number of examinees used in the study, the types of MST designs, are all considered for this selection. Finally 18 test forms were decided for the initial stage modules, and the test forms for the second stage and third stage modules were decided accordingly based on the inverse proportion method for every MST design. As in the MST 1-2 design, 18 parallel test forms are available to be selected and assembled at the initial stage, and nine forms are available to be assembled for the easy module and nine for the difficult module at the second stage.

The exposure rate for this study is 20% and the exposure control procedure is implemented at the module level. A randomization method is applied in the study to implement the exposure control. To implement this method, the items are first randomly selected within bins based on the number of items required for each module. By random selection, all the items

constructed for parallel modules share equivalent statistical attributes, and modules constructed also have no overlapping items across them. Examinees are then administered one of the multiple test forms at each stage. When one particular test form is administered to the examinee, and if the exposure rate is smaller than 0.2, this test form would be administered. Otherwise, it would be excluded from selection and the test form would be selected from the remaining ones. The above procedure continues until all the examinees complete the whole testing procedure for various MST panel designs.

3.1.4 Test assembly

In test assembly, one advantage of the bottom-up approach is that it addresses both statistical and non-statistical requirements of a test at the module level so that they can be mixed and matched easily at each stage to form parallel panels. Given this advantage, the bottom-up approach was applied in the present study for test assembly purposes.

In the simulated optimal item pools, the item difficulty distributions for the hypothetical tests at various stages are assumed the same. Based on this assumption, the assembling of panels could be completed by mixing and matching the forms of the modules, and the pathways are parallel between different assembled panels. The initial routing test has a range of medium item difficulty, and it is divided into easy and hard modules or easy, medium and hard modules at later stages for different test designs. The range of item difficulty across modules at the same stage is allowed to have an overlap because this benefits the examinees whose abilities are at the border of two modules. The TIFs target set up in the design stage is 20 per pathway with a reliability of 0.95 for the whole test. That is, the accumulated item information for all modules along each pathway is supposed to reach 20 to ensure the high reliability of the MST test constructed.

Panel assembly is completed in two steps: assembling items into modules from the item bank and assembling panels from the modules. A bottom-up approach was applied in the present study for test assembly purpose with content constraints controlled at the module level. In this way, the modules are parallel and the assembling of panels could be completed by mixing and matching the forms of the modules, and the pathways are parallel between different assembled panels.

3.2 Item pool design

With MST, the *p*-optimality method was used to determine the items needed for an optimal item pool and support a particular panel design. The basic idea was to randomly select examinees from a target population and simulate the MST procedure. With the application of the Rasch model in the study, the item difficulty parameter was the major factor to determine the bin size and item selection. If the item that is available for selection has information that is within 95% of the maximum possible, the bin width is identified as .35 (Reckase, 2003). That is to say, if the item difficulty is .35 logit away from ability estimate, the loss of information for examinees with the selected item is 5% compared with a perfect match to the ability estimate. Considering the small amount of information lost for an item, a bin width of .35 was applied in the present study. The selected items were sorted into a set of "bins" which were defined on examinees' proficiency scale. After a certain number of examinees took the simulated MST tests, the number of items in the item pool stabilized and reached an optimal quality, and the union of the items constituted the optimal item pool blueprint for that MST panel design.

Taking the 1-2 design as an example, the detailed procedures for applying the *p*-optimality method in the context of MST were:

1) Identify the mean of the true abilities from the target population (e.g., standard normal distribution) in the study and use it as the cutoff score for the routing test;
2) randomly select examinees from the target population;
3) simulate the administration of items matching the examinees' true abilities for the routing test in which items are selected from the bin containing the mean of the true abilities;
4) route the examinees to one of the two modules at the second stage using the cutoff score identified for routing;
5) simulate the administration of items matching the examinees' true abilities for the number of items in the routed module at stage 2;
6) stop the procedure when the anticipated measurement precision is acquired for this fixed-length test;
7) assign all items selected in both the routing test and stage 2 to the bins;
8) conduct 100 replications to reduce the sampling error,
9) obtain the average number of items required for each bin (e.g., rounded to the nearest whole number) and produce a bin count table, which indicates how many items are identified within each bin;
10) randomly select item parameters based on the bin count table from each bin to meet the test length requirement.

For the 1-3 design, step 1) is revised to include two cutoff scores based on 33^{th} percentile and 67^{th} percentile of the true ability distribution for the examinees. Step 3) is revised to include the administration of items matching the examinees' true abilities for the routing test in which

items are selected from the two bins containing the two cutoff scores. In step 4), the examinees are routed to one of the three modules at stage 2, then steps 5) to 10) are repeated.

For the 1-2-2 design, first steps 1) to 5) are repeated. In step 6), two cutoff scores are selected to route equal proportions of examinees to the final stage. In step 7), the procedure stops when the anticipated measurement precision is acquired for this fixed-length test. In step 8), assign all items selected in all stages to the bins. The rest of the steps from 9) to 11) are the same as 8) - 10) for MST 1-2 design. The two cutoff scores at stage 2 are calculated based on the observed score variance of the examinees with the following procedure. First, a fixed number of 5,000 examinees are selected from the target population and administered the routing test. Second, reliability is calculated based on the examinees' responses to all the items in the routing test. Third, the observed score variance is obtained with the known reliability and true ability variance (Thissen, 2000). Since the items in the first stage are selected from one bin and all examinees are exposed to the same set of items, marginal reliability is not used in this step. Fourth, the cutoff scores are determined based on the 25^{th} percentile and 75^{th} percentile of the observed score variance of the examinees. Fifth, based on the examinees' performance at stage 2 modules, they are routed to one of the two modules at the final stage. The procedures for 1-2-3 design are similar to those of the 1-2-2 design. They differ in the second stage where two cutoff scores are determined (e.g., the cutoff scores are determined based on the 33^{th} percentile and 67^{th} percentile of the observed score variance of the examinees) and the final stage where examinees are routed to one of the three modules represented by easy, medium and difficult items.

The set of items that are selected in this process constitutes the optimal item pool used for item selection for the various MST designs. Under the exposure control condition, the number of items within each bin is adjusted according to the number of items that are required by the

inverse proportional exposure control method. For example, at stage 1, 18 test forms are established to control item exposure rate for all MST test designs. The number of needed items is obtained through multiplying 18 and the proportions of routing test for various conditions. At stage 2, for 1-2, 1-2-2 and 1-2-3 designs, 9 forms are constructed for the easy module and 9 forms are constructed for the hard module. The numbers of needed items are obtained similarly with those at the first stage. For 1-3 design, 6 forms are constructed respectively for the easy, medium and hard modules. At stage 3, for 1-2-2 design, 9 forms are constructed for the easy module and 9 forms are constructed for the hard module. For 1-2-3 design, 6 forms are constructed respectively for the easy, medium and hard modules. All the selected items are combined together and constitute the optimal item pool under exposure control conditions. The item pool established this way for each design tailors to the distribution of examinees' true abilities.

Because the simulated test is a generalized test with no specific test blueprint and unidimensionality was assumed in the study, content balancing is not considered in the study. Assuming unidimensionality, item distributions for all hypothetical content areas are assumed the same for the simulated optimal item pool. In previous studies where the *p*-optimality method was applied in CAT (e.g., He & Reckase, 2014), unidimensionality was also assumed, but content balancing was considered. The results showed that item distributions across all the content strands were similar and the item exposure rate distribution for each content strand also shared a similar pattern. Another study by Zhou (2012) focused on the impact of exposure control and content balancing on the optimal item pool design for a polytomous CAT using the *p*-optimality method. The results obtained indicated that the practical constraints of content balancing had little effect on the item pool size and little impact on the item pool design for the

polytomous items. In the conditions where only the content balancing constraint differs, the distributions of item parameters were quite similar.

3.3 Research design

This section is composed of two parts. The first part introduces the simulation study design, which investigates the effectiveness of applying the *p*-optimality method to design optimal item pools and support the various MST designs. It also studies the impact of test length and routing test length as well as the exposure control on measurement accuracy. The second part investigates whether the *p*-optimality method is applicable to support the optimal item pool design for an operational MST.

3.3.1 Simulation study

A simulation study was first conducted to design the optimal item pools and support the different MST designs for with and without exposure control conditions. To achieve this end, examinees were first randomly selected from a standard normal distribution. These examinees were then administered the simulated test items following the various MST panel designs (e.g., 1-2, 1-3, 1-2-2, and 1-2-3) and routing rules discussed above to build up the correspondent tests. The maximum likelihood estimation method (MLE) was applied to estimate examinees' abilities. One hundred replications were conducted in this step to reduce the sampling error. Items required by these examinees were placed into a set of bins ranging from -3.5 to 3.5 with a bin width of 0.35.

After the number of items was obtained within each bin, specific item parameters were randomly selected from the bins assuming a uniform distribution. Thirty-six (4 test designs * 3 test lengths * 3 routing proportions) item pools were designed using the *p*-optimality method to support the different MST panel designs. Exposure control was not implemented at this step. To

examine how the exposure control affects the item pool size, for each condition, two item pools were compared: one with the exposure control and one without. The final number of item pools constructed was 72. The randomization method discussed above was used to control the item exposure rate, which was set as 20% in the study, a commonly used cutoff value in MST. The number of forms assembled for each module was inversely proportional to the number of modules of the stage it belonged to and an equal proportion of examinees were routed to different modules at later stages. Thirty replications were conducted for each condition to reduce sampling error. All programming work in this study was implemented by MATLAB 7.10.0 (R2010a).

The evaluation of the item pool performance was carried out by using overall and selected samples respectively. A simulated fixed sample of 5,000 examinees was randomly drawn from a standard normal distribution and were administered the four MST tests using all the candidate item pools. The procedures were also conducted with selected samples. The performance for all candidate item pools were evaluated by the conditional samples, using simulated examinee samples with proficiency points equally spaced over the range between -3.5 and 3.5 at an interval of .05. At each proficiency point, a simulated fixed sample of 100 examinees was drawn and were administered the four MST tests using all the candidate item pools.

Evaluation criteria include: Overall and conditional statistic values. The overall statistic values include: correlation between true abilities and estimated abilities, overall bias, root mean squared error (RMSE), test information, standard error of measurement, marginal reliability, item overlap rate and item exposure rate, and classification accuracy. The conditional statistic values include: conditional bias, root mean squared error (RMSE), standard error of

measurement, item overlap rate, and classification accuracy. Three different types of cutoff scores for median value, minimum competent and scholarship decisions are unitized for classification accuracy.

3.3.2 Application of the *p*-optimality method in an operational MST context

In this empirical study, the *p*-optimality method was applied to design an optimal item pool to support an operational licensure MST and see if it is applicable in real educational settings. The operational MST has a 1-2-2 design and all the items come from one content area in the test. The test has 75 operational items, and the item parameters were calibrated from a total sample of 27,261 examinees using Rasch model by fixing the person parameters in WINSTEPS. Based on the probit model scale, the calibrated item parameters have the mean of 0.97 and the variance of 0.59. The calibrated operational item pool consists of 1029 items.

Following the procedures discussed in the methodology section, a group of 5,000 examinees were selected from a standard normal distribution and administered the test assembled from the real pool built with the *p*-optimality method. Since the real pool contained more items than needed, an item pool assembly process was involved. In this process, the desired number of items for each module was selected based on the test specification (e.g., 25 items per module) of the operational test. The operational item pool had the capacity of assembling eight test forms for each module. Since the master item pool only contained items from one content area of the subject, content balancing was not considered in the item pool assembly process. The number of items within each bin was determined following the procedure discussed above in the methodology section. After the bin counts or the number of items within each bin were established, the real item pool parameters were re-distributed and filled out the bins as required by the bin counts. The width of the bin is still .35. For the real item pool, if there were no enough

items within certain bin, the adjacent items that were closest to them in terms of item difficulty would be moved to that bin. Since the target population had a larger variance as compared to the item parameter distribution in the real pool, and the bins covered a wider range along the θ scale, almost all the items in the real pool needed to be re-distributed to new bins. After this re-distribution, although the item parameters were not aligned perfectly with the boundaries of bins on the proficiency scale, they were as close as possible to the targets. In this process, 29 redundant items close to the middle range of the ability scale were removed from the pool. The re-distributed real item pool is called the *R*-Pool.

As one of the anticipations in the *p*-optimality method is the alignment between target examinee population and the item parameter distribution in the optimal item pool, a simulated optimal item pool was created to compare with the real item pool performance. This simulated item pool is called the *S*-Pool. The examinee population used to design the optimal item pool was the same with that of the real pool situation, which also came from a standard normal distribution. The same group of 5,000 examinees (e.g., same with that in the operational item pool situation) was used to design the simulated item pool following the procedures as discussed in the methodology section. To enable the comparability between the item pool performance between the *R*-Pool and *S*-Pool, the simulated test for both followed the test configurations of the operational test (e.g., 75 item test with 1-2-2 MST design, and 25 items per module), exposure control procedure and content balancing. To implement the exposure control, all examinees were exposed to an equal number of modules at each stage. For example, at stage 1, each examinee was exposed to eight test forms, and after being routed to one of the two modules at stage 2, he/she was still exposed to eight test forms. The number of test forms built also followed that used in the operational test. Since the operational test comes from one section of the licensure

test and all the items share one content area, content balancing was not needed in this process. The *p*-optimality method was applied to both the simulated and operational item pools to support the 1-2-2- MST design and 30 replications were conducted to reduce sampling error. The item parameter distributions and item pool sizes in the different stages of both the simulated and operational item pools were compared.

3.3.3 Study design elements

To sum up, for the simulation study, the research design involves the following factors: MST panel designs with four levels (e.g., 1-2, 1-3, 1-2-2, 1-2-3); test lengths with three levels (e.g., 20, 40, and 60); routing test proportion with three levels at the first stage for all designs (e.g., 20%, 30%, and 40%), three levels at the second stage for MST 1-2 and MST 1-3 designs (e.g., 80%, 70% and 60%), three levels at the second stage for MST 1-2-2- and MST 1-2-3 designs (e.g., 40%, 30%, 20%), and one level at the third stage for MST 1-2-2 and MST 1-2-3 designs (e.g., 40%); exposure control with two levels (e.g., without exposure control and with exposure control). For the empirical study, the research design involves the following factors: MST panel design with one level (e.g., 1-2-2); test length with one level (e.g., 75-item test); module length with one level (e.g., 25 items per module); exposure control with two levels (e.g., without and with exposure control); application of the bin-and union method with two levels (e.g., real item pool and simulated item pool).

3.3.4 Evaluation criteria

The following section provides details or equations of how the overall evaluation statistics and conditional evaluation statistics are computed to evaluate the performance of the candidate item pools in the simulation study and the empirical study:

1) Equations used to calculate the correlation coefficients between estimated and true person abilities, overall and conditional bias, root mean squared error (RMSE), overall and conditional item overlap rate are given by the following equations:

$$r_{\theta,\hat{\theta}} = \frac{\sum_{j=1}^{N}(\hat{\theta}_j - \theta_j)(\hat{\theta}_j - \bar{\hat{\theta}})}{S_\theta S_{\hat{\theta}}} \tag{10}$$

$$Bias = \frac{1}{N}\sum_{j=1}^{N}(\hat{\theta}_j - \theta_j) \tag{11}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{j=1}^{N}(\hat{\theta}_j - \theta_j)^2} \tag{12}$$

where $\hat{\theta}_j$ and θ_j are the estimated and true abilities of the j^{th} examinee. S_θ and $S_{\hat{\theta}}$ are the standard deviations of the true and estimated abilities. *N* is the number of examinees.

2) Classification accuracy of the ability estimates. The classification accuracy is measured by dividing the number of examinees correctly classified by the total number of examinees. Three threshold scores are used in the study: classification accuracy based on the median value, 80% and 95% of the examinees' score distribution. The cutoff score based on the median value is used to evaluate the pass/fail of the test, and the one based on 80% is used to evaluate whether the examinees possess a minimum competence in the subject area of the test, and the one based on 95% is used to evaluate if the examinees deserve any scholarship. Conditional classification accuracy is calculated using the number of examinees correctly classified conditional on the ability points from -3.5 and 3.5 at an interval of .05.

3) Item exposure rate. It is obtained through dividing the number of times the item is administered in the item pool by the number of examinees. For example, if the exposure rate is 0.1, it means the module is administered to 10% of the examinees. If the module exposure rate is too high, it means the set of items is being exposed to too many examinees and the test security will be threatened. The item exposure rate is considered high if it is over .20 (Segall et al., 1997). On the contrary, if the module is rarely used in test administration, it is called underexposed. Underexposure of items indicates the underutilization of the item pool. In CAT, an item with an exposure rate lower than .02 is considered as underexposed (He & Reckase, 2014). This value is adopted in the current MST context for evaluating the underexposure rate of modules.

To evaluate whether the items within bins at some parts of the θ scale are exposed more than the others, the item exposure rate conditional on bins is reported. Two evaluation statistic values are calculated: one is the average item exposure rate across bins, and another is the proportion of items above or below the threshold exposure rate as discussed above (e.g., <.02 or >.20). The procedures for calculating the two statistic values are as follows. First, after obtaining the item exposure rates for all items by the overall sample, sort them into different bins based on their item parameters. Second, calculate and report the average value of the item exposure rates across bins. Third, calculate and report the proportion of item exposure rate above the threshold value of .20 for each bin under no exposure control condition, and the proportions of item exposure rate above the threshold value of .20 as well as below .02 for each bin under exposure control condition.

4) Item overlap rate. It is defined as the number of common items shared by two randomly selected examinees divided by the test length in the test (Way, 1998). Both the average

item overlap rate and the conditional average item overlap rate were calculated. The following equation was used to calculate the average item overlap rate in the present study:

$$R=\frac{T/{C_N}^2}{\sum_{j=1}^{N} L_j/N}, \qquad (13)$$

where *T* is the total number of items shared by the given number of pairs in the tests among *N* examinees. L_j is the total number of items administered for *N* examinees.

The overlap across modules in terms of the item difficulty range is checked at two stages, one is at the item pool design stage and another is at the simulated item pool stage. The former is based on the boundaries of the bins for different modules, which are equivalent to the minimum and maximum item difficulty parameters that are aligned with the lower and upper bound of the bin for each module involved. This applies to both types of exposure control conditions. The latter is decided by the minimum and maximum values of the item parameters in the simulated item pool. Under exposure control condition, the minimum and maximum values are decided after evaluating all the items in the simulated pool.

CHAPTER 4: Results

This chapter presents the results from the study. It first describes the results of the optimal item pools designed by the *p*-optimality method to support the various MST designs. Then it shows results from the simulation study and the empirical study. The results from the simulation study include the evaluation of the MST design, test configuration, and simulated item pool performance. The results from the empirical study include the evaluation of the *p*-optimality method in an operational MST context through comparing it with a simulated optimal item pool.

4.1 Results from the optimal item pools designed by the *p*-optimality method

Figure 4.1 to Figure 4.8 showed the bin-counts for all the test configurations of all MST designs with and without exposure control. In the figures, n20, n40 and n60 mean the test length, and 20%, 30% and 40% mean the routing test proportions. For example, n20.20% means the condition of a test length of 20 and routing test proportion of 20%. In these figures, the different modules at the second stage or third stage were marked with different colors. In 1-2-2 design and 1-2-3 design, the two modules at the second stage were marked with the same color because they were always symmetric and contained the same number of items.

In each figure, the number of items within each bin for all the stages in the particular MST design was displayed. It is seen from the figures that the item frequency distributions across different MST designs and test configurations did not have any uniform characteristics. Generally speaking, some item distributions were symmetric and some were skewed. More items were distributed at the value of the cutoff score especially when the routing test proportions were large, such as 40%. The item distributions became more peaked with the increase of the routing test proportions. With the increase of the routing test proportions, the number of items

accumulated at the first stage became larger, and the items at the second and third stages in the various MST designs became comparatively fewer. From the observations we could see that the distributions of the item parameters for all the MST designs for the conditions of without exposure control and with exposure control were similar. Some differences are that the item pool sizes for under the exposure control conditions were much larger than those under no exposure control, and there were more items accumulated at the initial stages due to the implementation of the exposure control method.

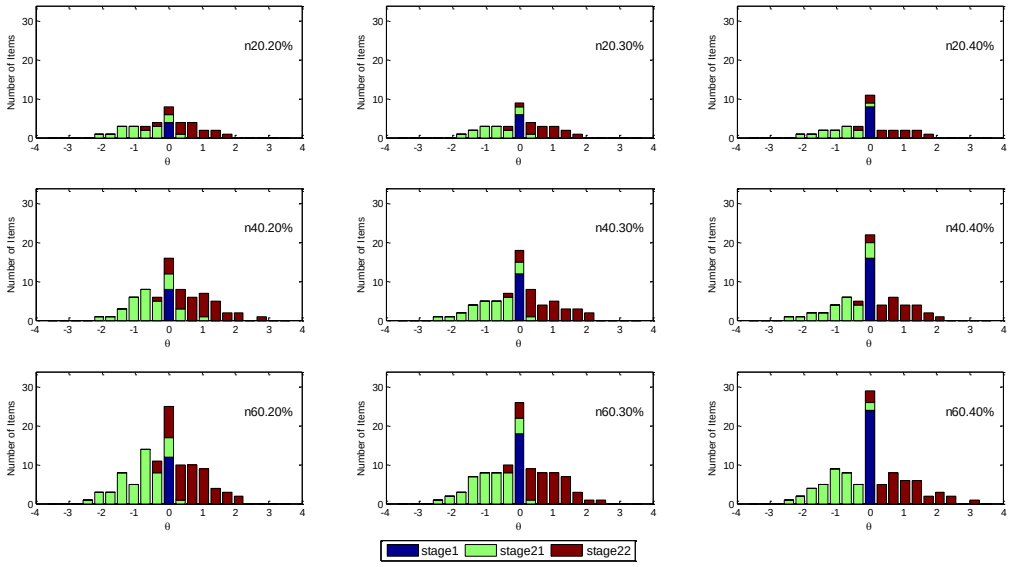

Figure 4.1 Number of items within bins for MST 1-2 designs without exposure control

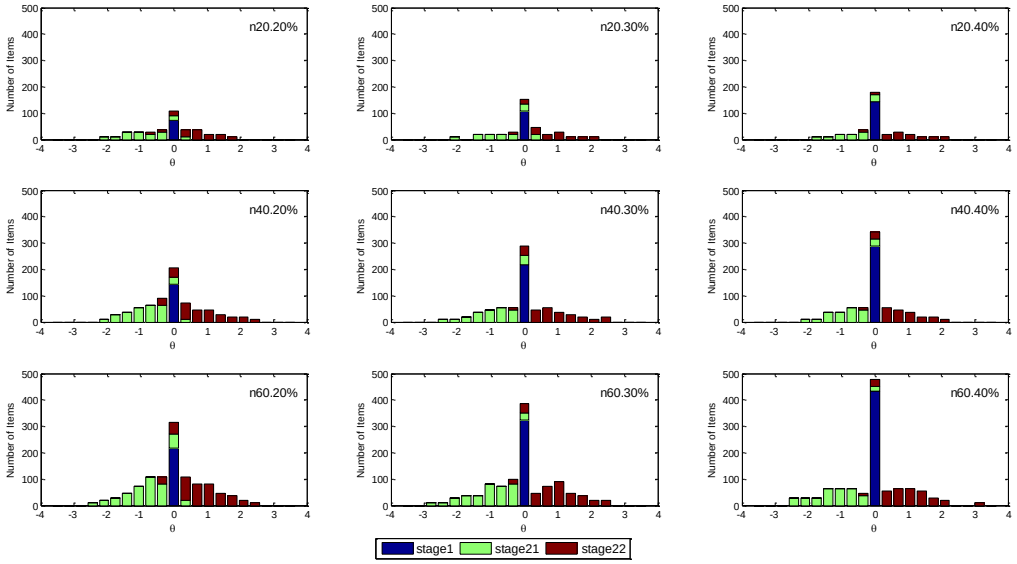

Figure 4.2 Number of items within bins for MST 1-2 designs with exposure control

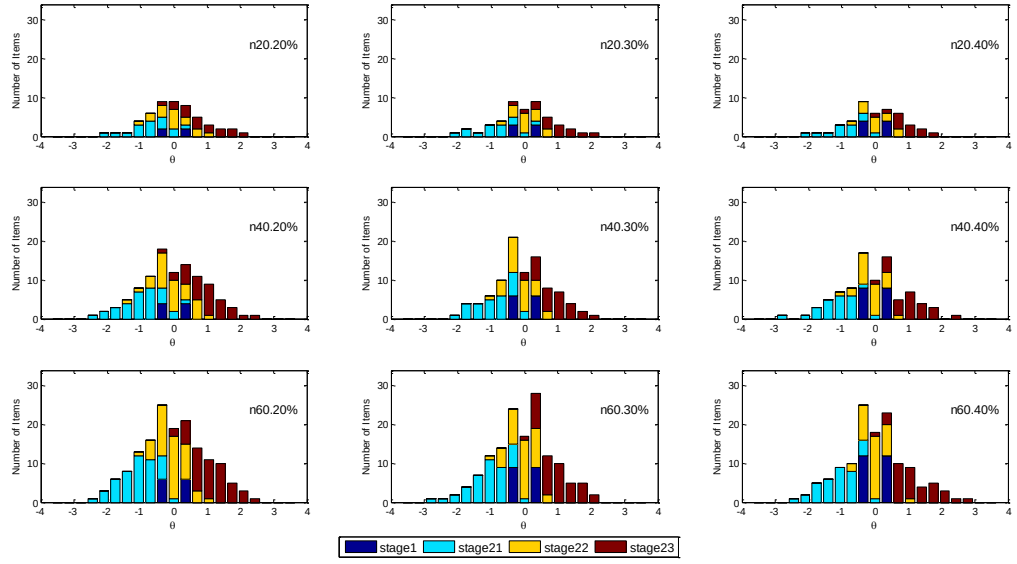

Figure 4.3 Number of items within bins for MST 1-3 designs without exposure control

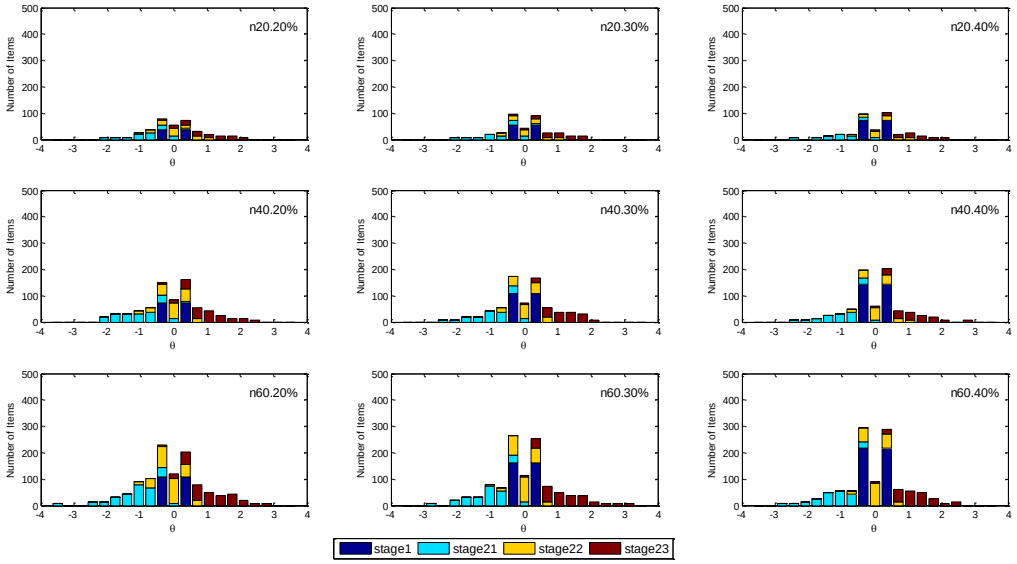

Figure 4.4 Number of items within bins for MST 1-3 designs with exposure control

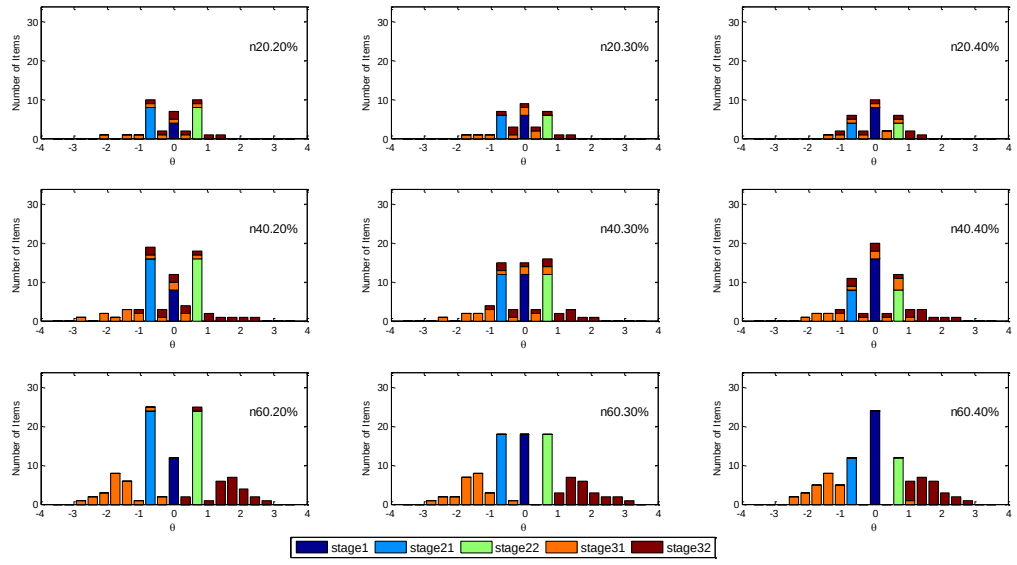

Figure 4.5 Number of items within bins for MST 1-2-2 designs without exposure control

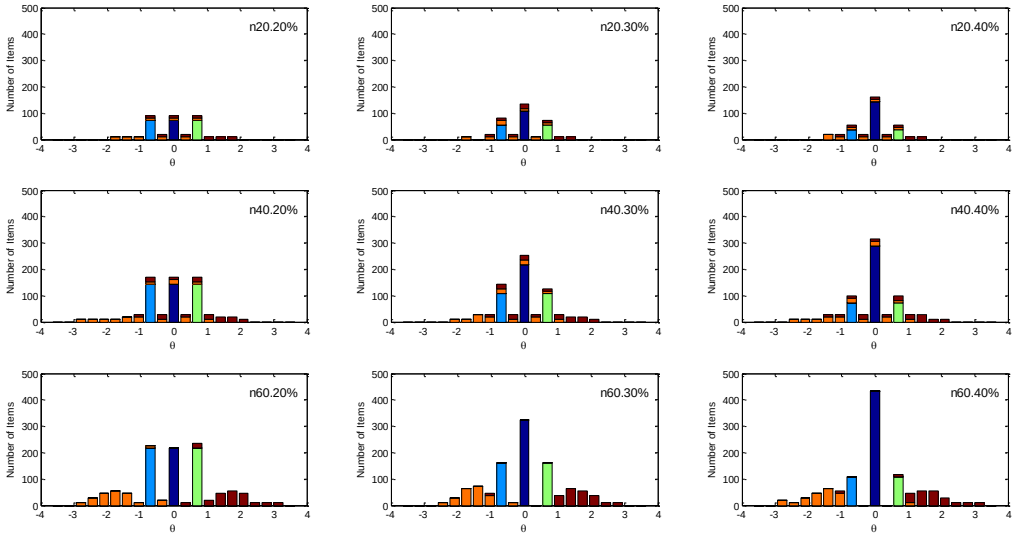

Figure 4.6 Number of items within bins for MST 1-2-2 designs with exposure control

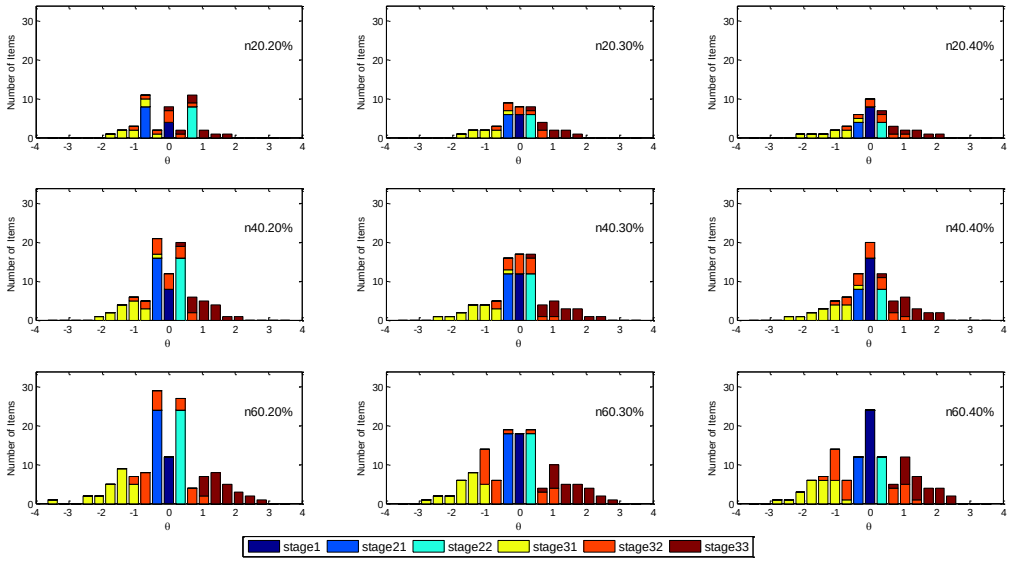

Figure 4.7 Number of items within bins for MST 1-2-3 designs without exposure control

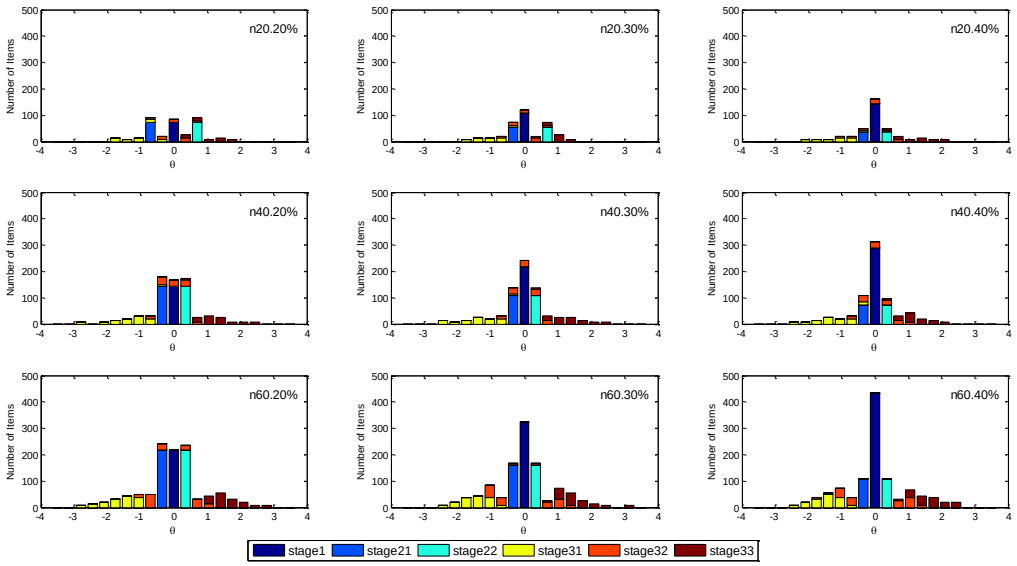

Figure 4.8 Number of items within bins for MST 1-2-3 designs with exposure control

Since there were a clear cut for the boundaries of item parameters at the first stage for all the designs, and at the second stage for 1-2-2 and 1-2-3 designs, the overlap across modules in terms of the item parameter range was only checked for the final stage modules in the item pool design stage. One test length of 40 with 30% routing test proportion was selected as an example to demonstrate this phenomenon. Figure 4.9 to figure 4.12 displayed the overlap of the range of the item parameters under no exposure control conditions. In the figures, the x-axis represents the examinees' ability, and the y-axis represents the probability. The curve indicates the probability density function and the lines display the item difficulty ranges for the different modules at the final stage in various MST panel designs.

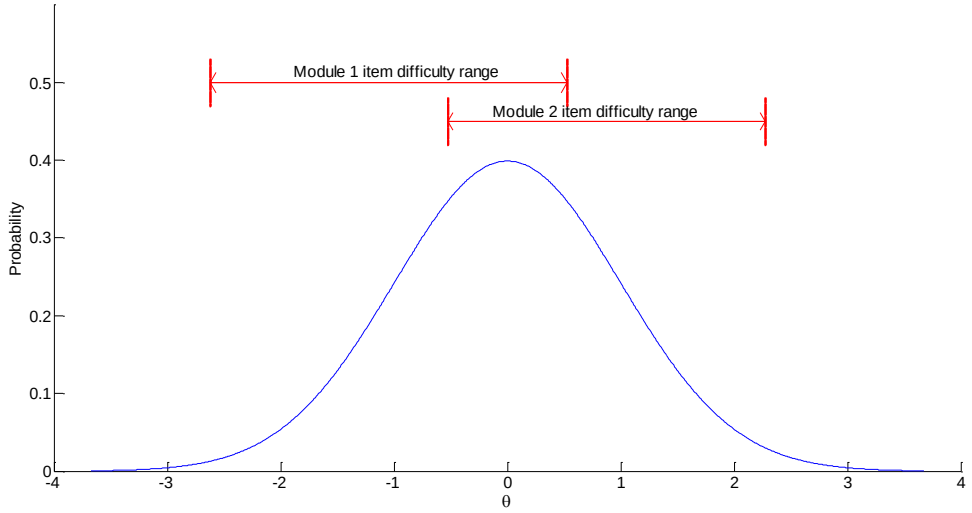

Figure 4.9 Item overlap across modules at stage 2 in MST 1-2 design without exposure control at item pool design stage

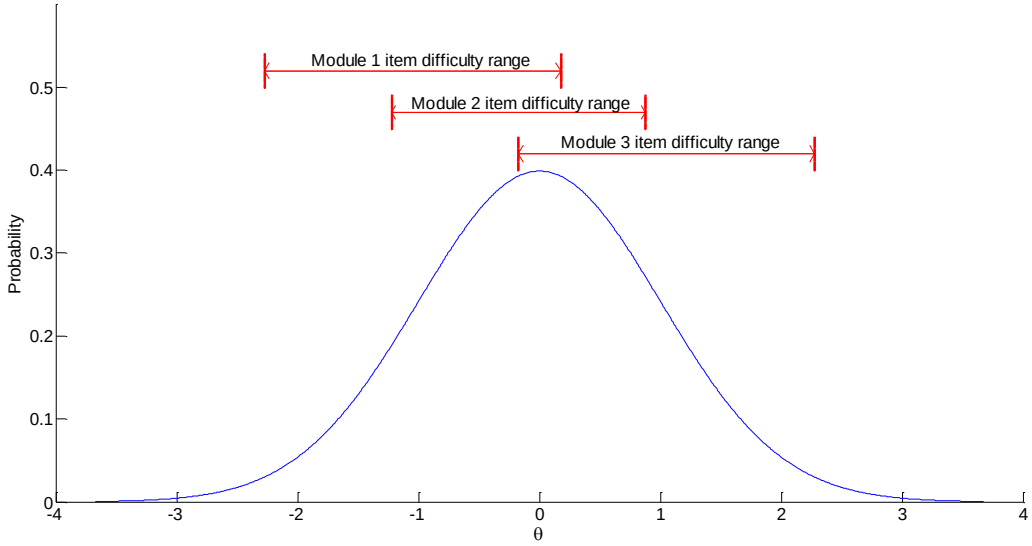

Figure 4.10 Item overlap across modules at stage 2 in MST 1-3 design without exposure control at item pool design stage

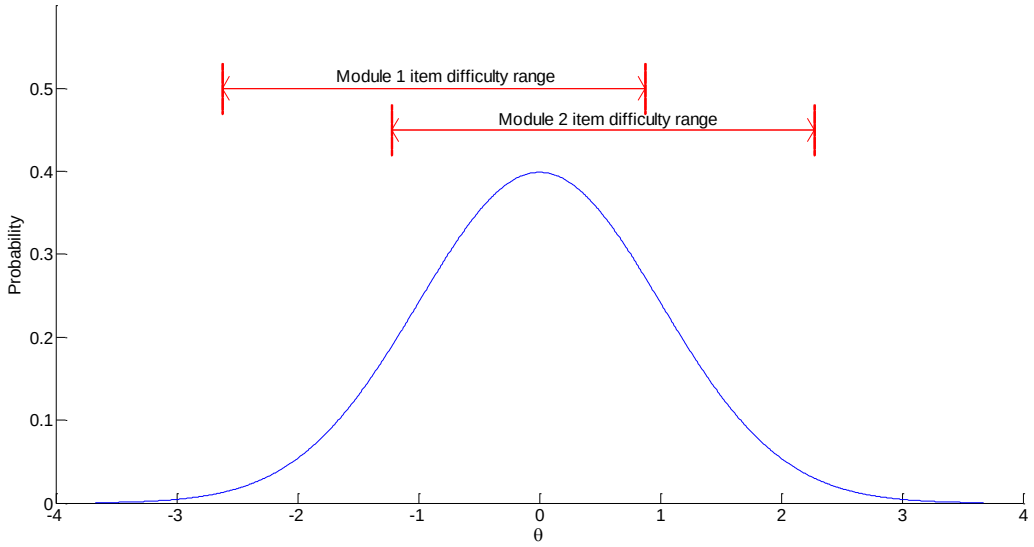

Figure 4.11 Item overlap across modules at stage 3 in MST 1-2-2 design without exposure control at item pool design stage

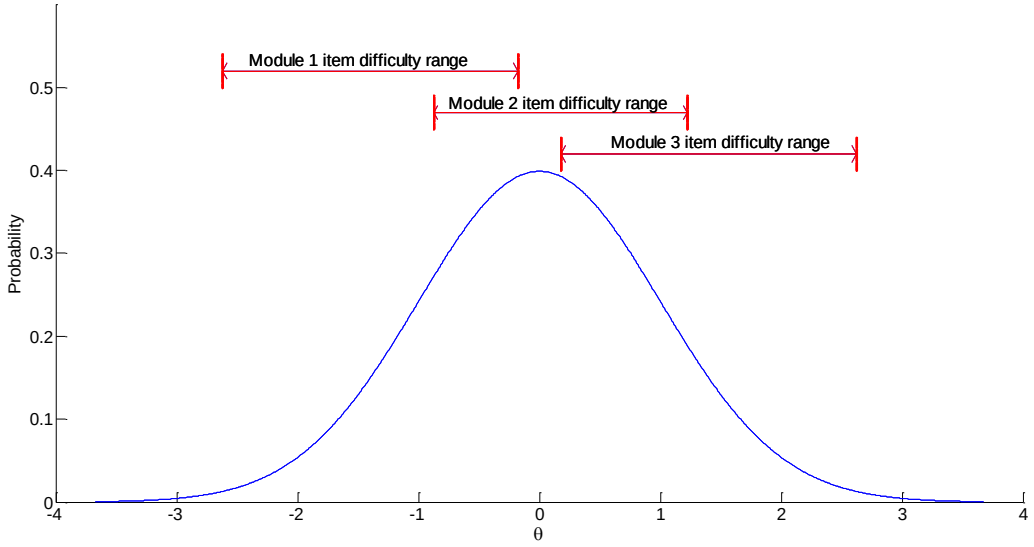

Figure 4.12 Item overlap across modules at stage 3 in MST 1-2-3 design without exposure control at item pool design stage

Figure 4.13 to Figure 4.16 display the overlap of the range of the item parameters under exposure control conditions at the item pool design stage.

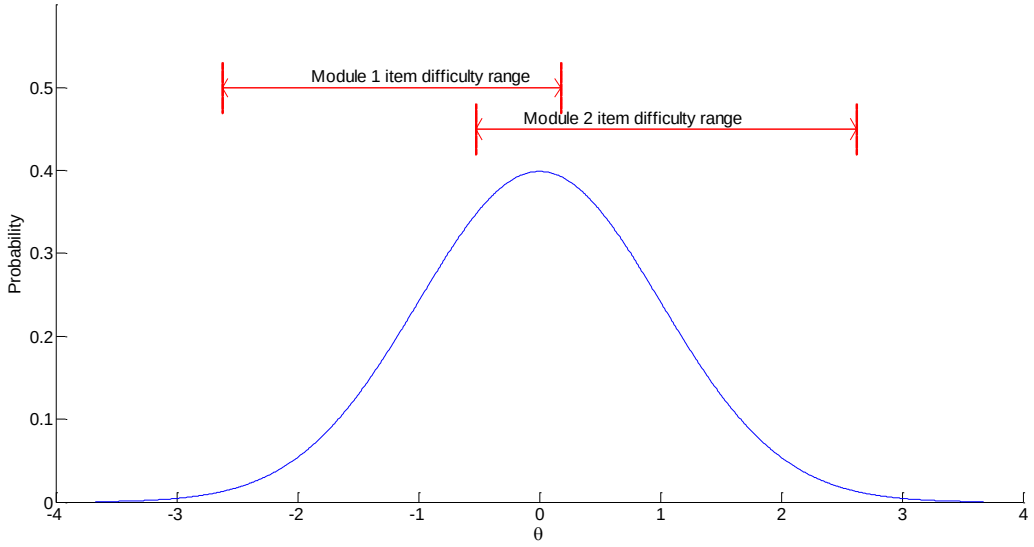

Figure 4.13 Item overlap across modules at stage 2 in MST 1-2 design with exposure control at item pool design stage

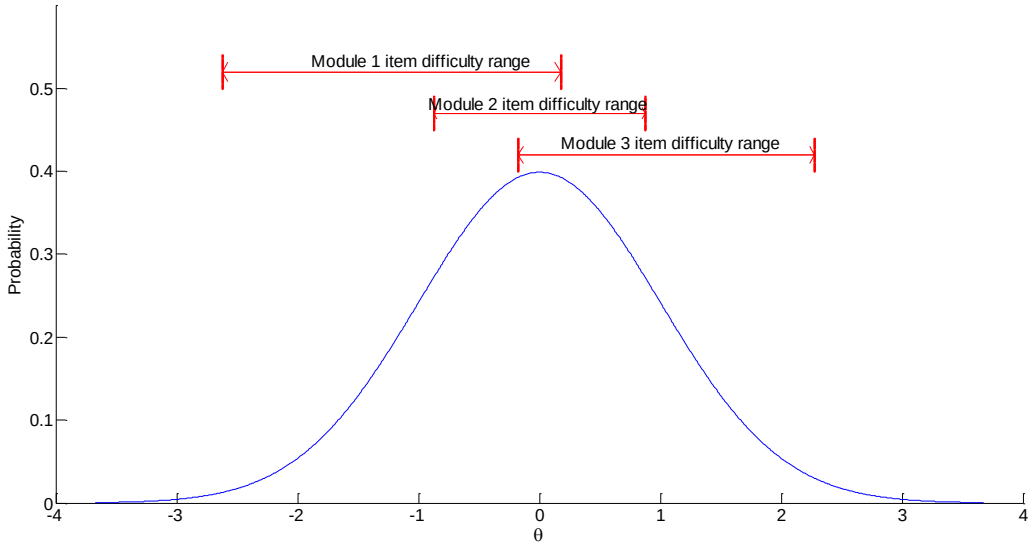

Figure 4.14 Item overlap across modules at stage 2 in MST 1-3 design with exposure control at item pool design stage

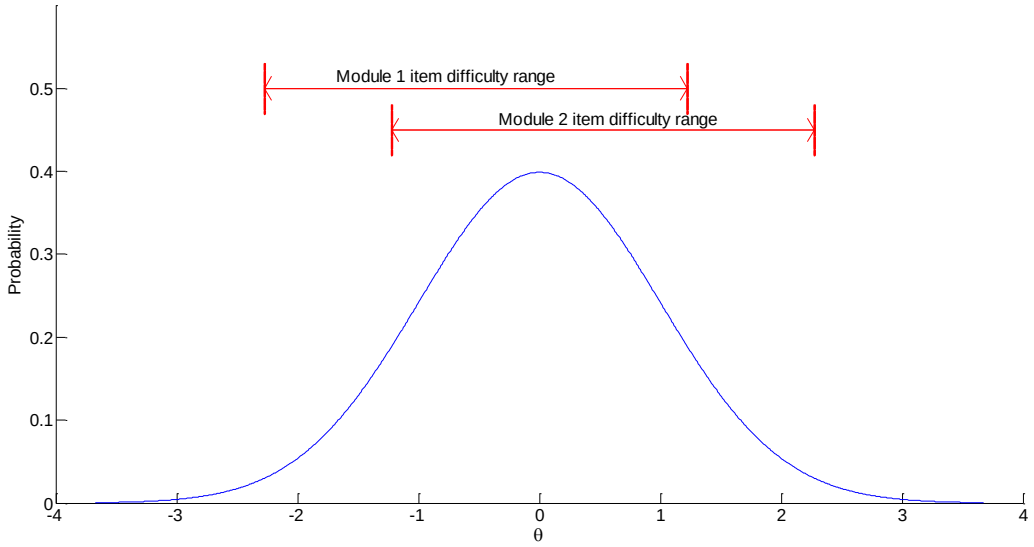

Figure 4.15 Item overlap across modules at stage 2 in MST 1-2-2 design with exposure control at item pool design stage

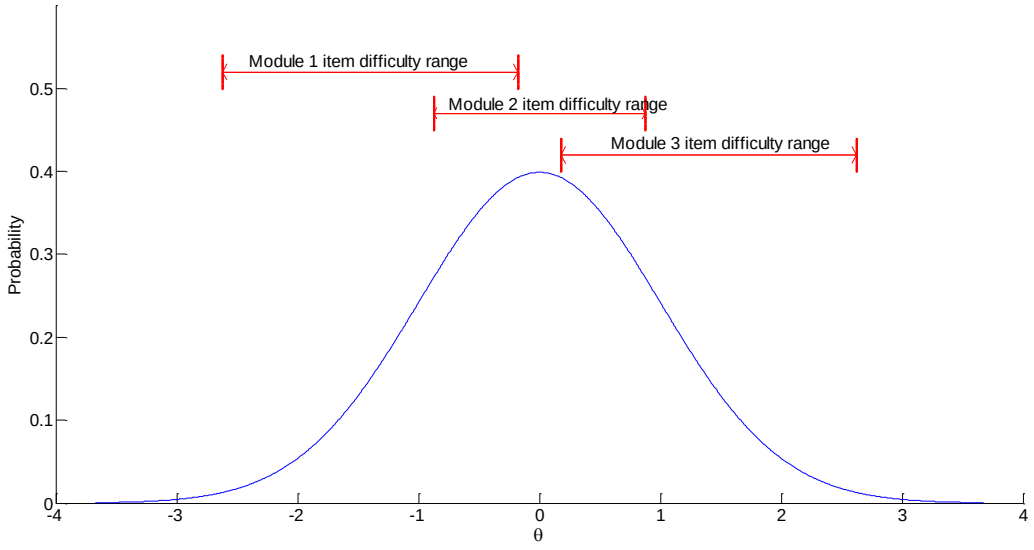

Figure 4.16 Item overlap across modules at stage 2 in MST 1-2-3 design with exposure control at item pool design stage

There was an overlap across modules in terms of the item difficulty range for all the designs under both exposure control and non-exposure control conditions. However, no obvious differences were discovered from the comparisons between the two sets of figures under these two exposure control conditions. In order to check whether the developed tests were aligned with the item pool design features, the overlap across modules in terms of the item parameter range was also checked in the simulated item pools. One test length of 40 with 30% routing test proportion was selected as an example to demonstrate this phenomenon. Figure 4.17 to Figure 4.20 display the overlap of the range of item parameters across modules under no exposure control conditions.

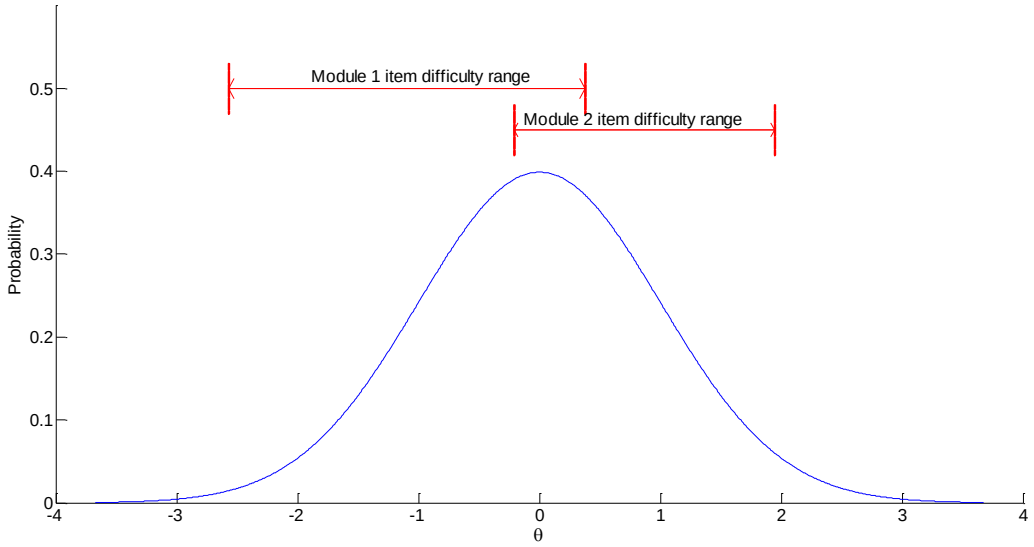

Figure 4.17 Item overlap across modules at stage 2 in MST 1-2 design without exposure control in the simulated item pool

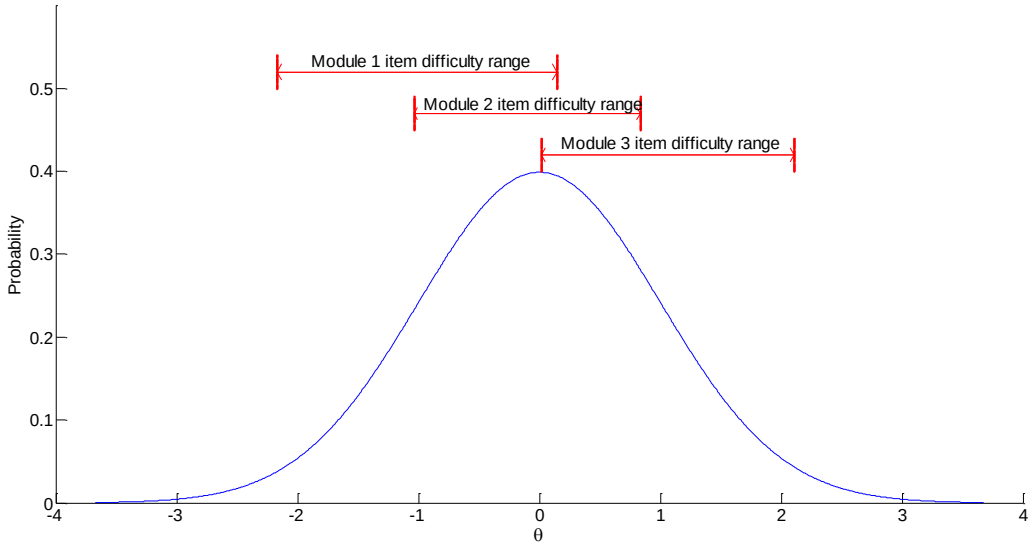

Figure 4.18 Item overlap across modules at stage 2 in MST 1-3 design without exposure control in the simulated item pool

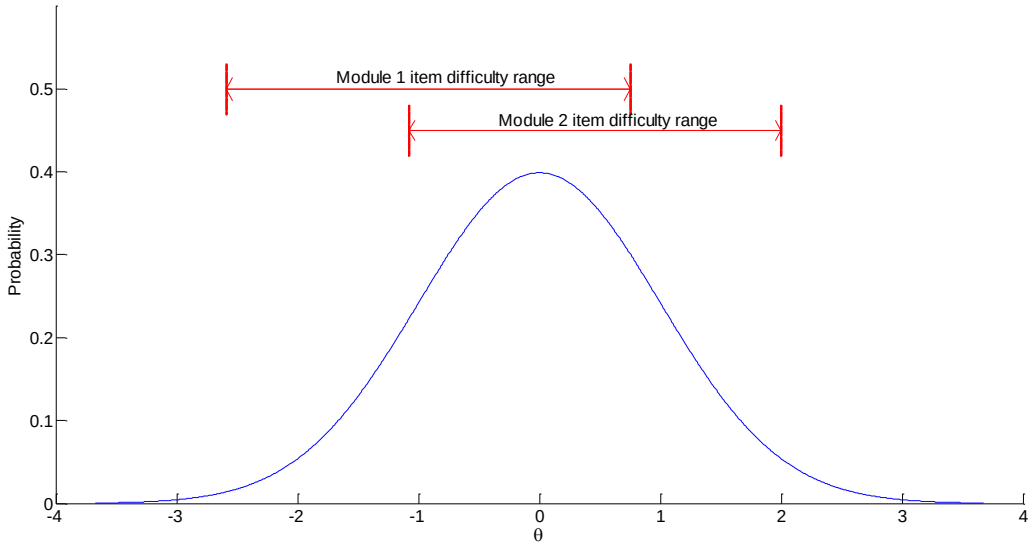

Figure 4.19 Item overlap across modules at stage 2 in MST 1-2-2 design without exposure control in the simulated item pool

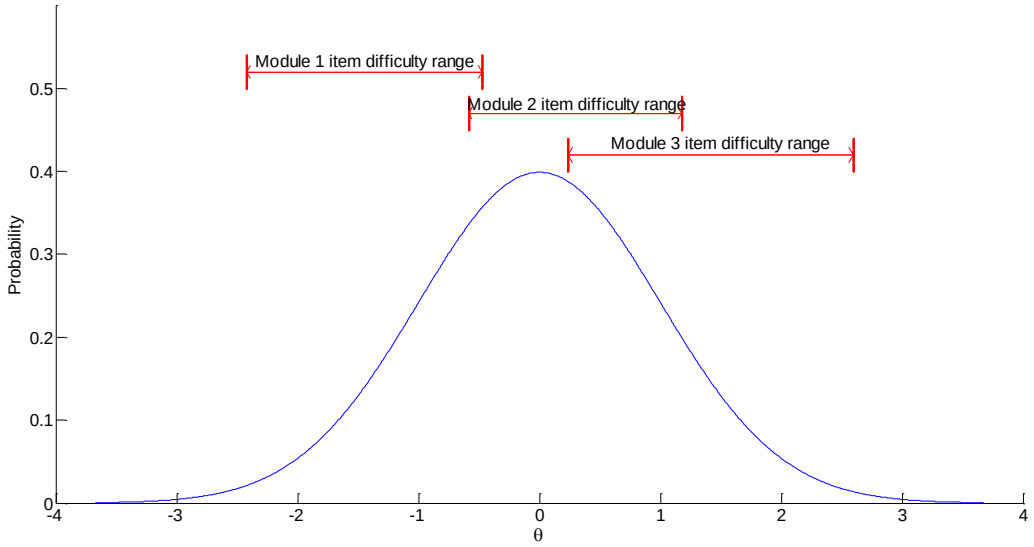

Figure 4.20 Item overlap across modules at stage 2 in MST 1-2-3 design without exposure control in the simulated item pool

Figure 4.21 to Figure 4.24 display the overlap of the range of the item parameters under exposure control conditions for the simulated item pools.

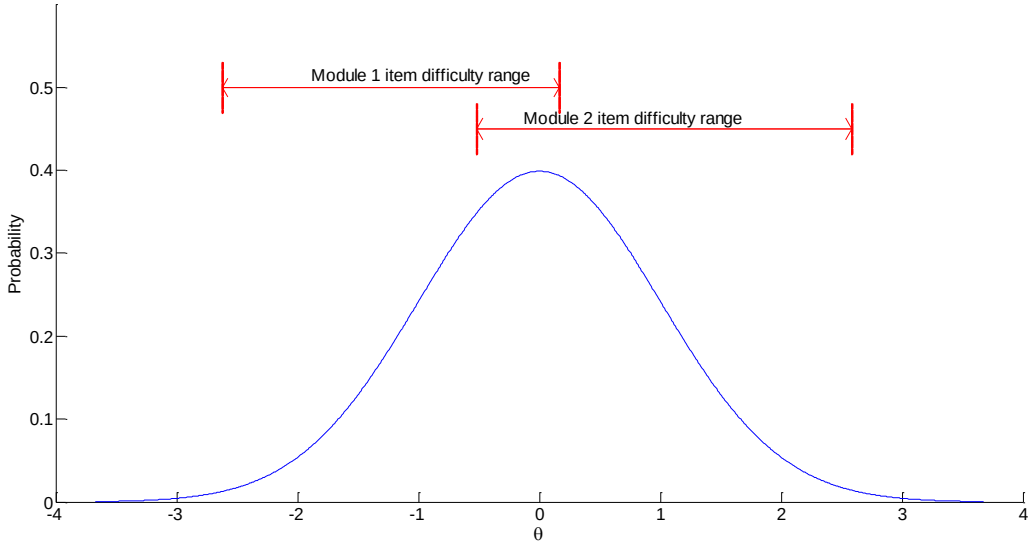

Figure 4.21 Item overlap across modules at stage 2 in MST 1-2 design with exposure control in the simulated item pool

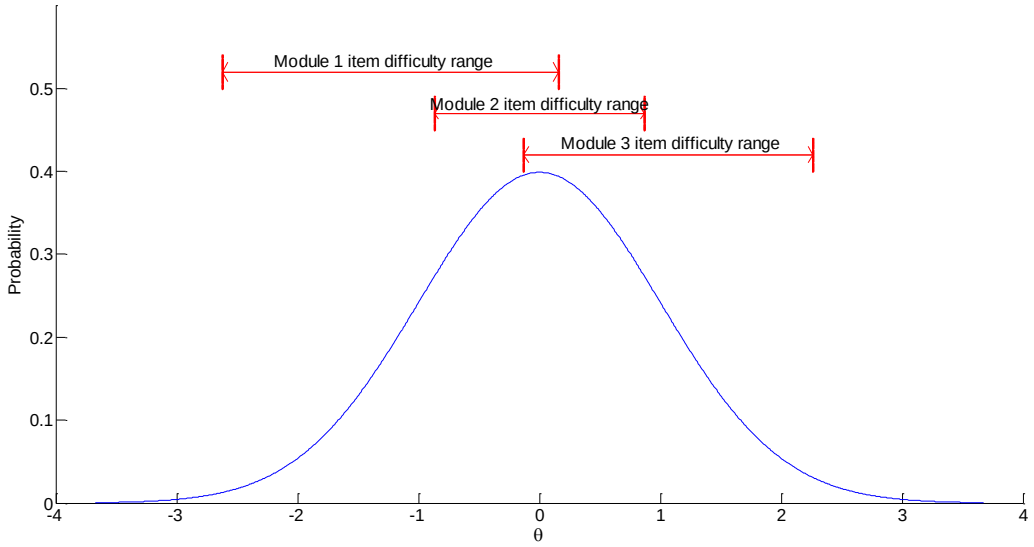

Figure 4.22 Item overlap across modules at stage 2 in MST 1-3 design with exposure control in the simulated item pool

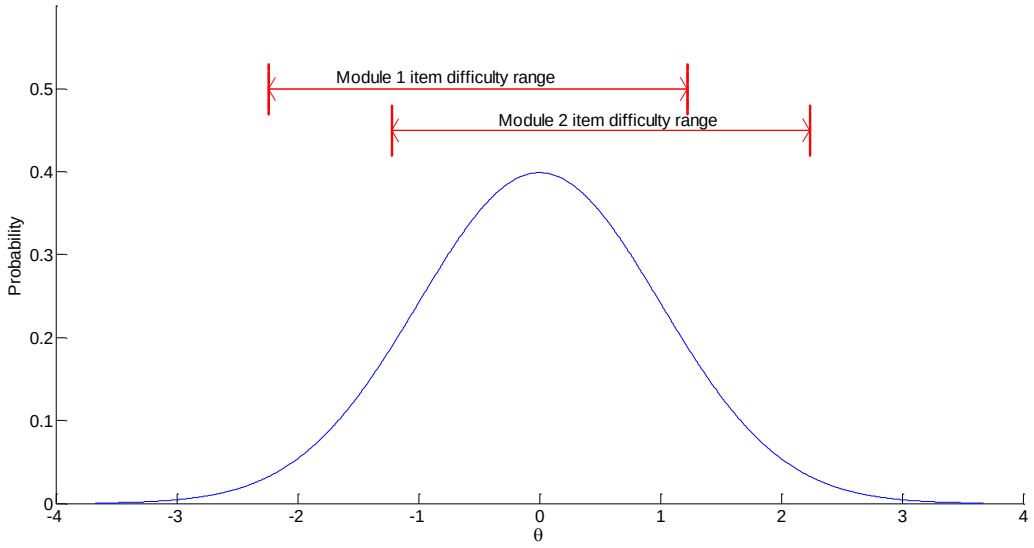

Figure 4.23 Item overlap across modules at stage 2 in MST 1-2-2 design with exposure control in the simulated item pool

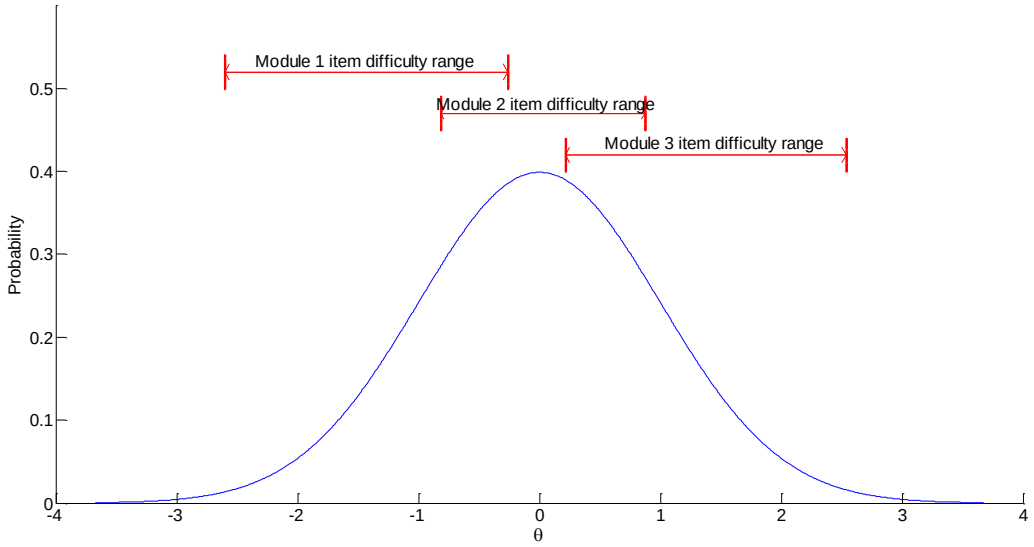

Figure 4.24 Item overlap across modules at stage 2 in MST 1-2-3 design with exposure control in the simulated item pool

The comparison between the results of the overlap across modules in the simulated item pool with those at the item pool design stage showed that the items developed in the simulated pool met the requirement of the item pool design.

The test information functions in the test development stage for all MST designs were shown in Figure 4.25. The vertical comparison is for different test lengths, and the horizontal comparison is for different MST designs. Within each figure, the results from different routing test proportions are compared. The test information functions were obtained by taking the average information for all the 100 replications in the test design stage. This curve was computed by setting up intervals on the θ scale and calculates the test information for each θ level. This procedure was replicated 100 times and the average values were used to draw the test information function curves. Except for some test configurations in MST 1-2 and MST 1-2-2 designs where a bimodal distribution was identified, most of the information distributions were normal, especially for the ones with 40% routing test proportions. The upper-stage test provided

similar levels of test information as the lower-stage tests for all the designs. Except for the test lengths of 20, the rest of the tests provided more than the required test information over the range from -1 to 1 on the θ scale. These figures applied to the situations for both with and without exposure control.

Figure 4.26 shows the module information curves for all MST designs under all test configurations for the test length of 40 with no exposure control. MST12.n40.20% means the condition of MST 1-2 design with a test length of 40 and routing test proportion of 20%. M means module. For all the MST design, M1 means the module at Stage 1. For MST 1-2 design, M21 means the easy module, and M22 means the hard module at Stage 2; for MST 1-3 design, M21 means the easy module, M22 means the module with medium difficulty, and M23 means the hard module at Stage 2; for MST 1-2-2 design, M21 means the easy module, M22 means the module with medium difficulty at Stage 2, M31 means the easy module, and M32 means the hard module at Stage 3; for MST 1-2-3 design, M21 means the easy module, M22 means the module with medium difficulty at Stage 2, M31 means the easy module, M32 means the module with medium difficulty at Stage 3, and M33 means the hard module at Stage 3. It was observed that the information curves for the modules at the same stage were almost symmetric. The module information curves for the first stage were more peaked for 1-2-2 and 1-2-3 designs and when the tests were longer. The module information curves for the test length of 20 and 60 are displayed in the Appendix (See Figure A.1 and Figure A.2). The test information functions and module information functions in the following figures all showed that the item pool parameters selected using the *p*-optimality method were appropriate to support the different MST panel designs.

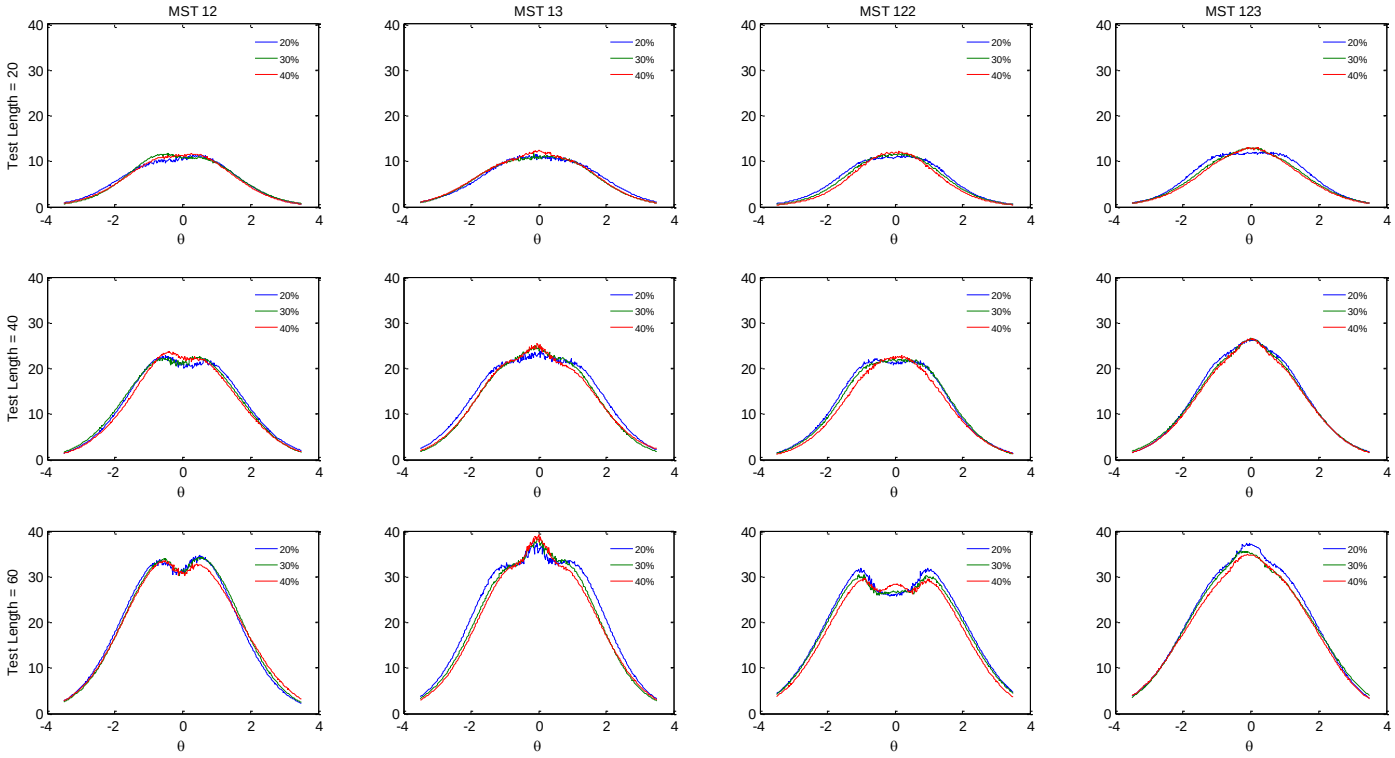

Figure 4.25 Test information functions for all test configurations in all MST designs at the item pool design stage

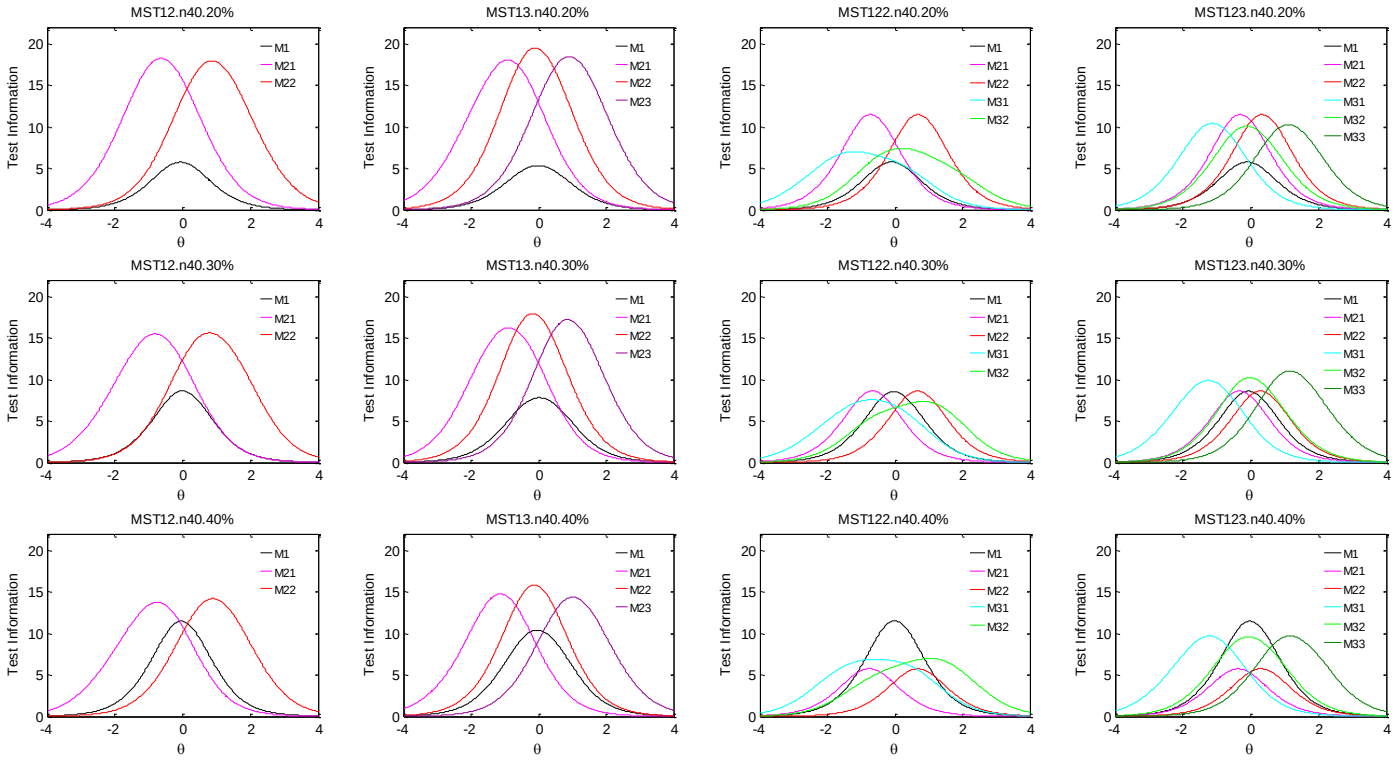

Figure 4.26 Module information curves for all test configurations in all MST designs at the test length of 40

4.2 Results from the different MST designs

These results come from the simulation study, which are summarized in two parts. The first part shows the evaluation of the optimal item pool performance to support the various MST designs through investigating the evaluation results within MST designs and across exposure control conditions. The second part shows the comparison between different MST designs through comparing the overall and conditional evaluation statistics.

4.2.1 Results within the different MST designs

Table 4.1 to Table 4.8 displayed the descriptive statistics for the item pools of all the MST designs under the conditions of without exposure control and with exposure control.

Table 4.1 Item pool descriptive stastitics of the MST 1-2 design without exposure control

Routing Proportion	Pool Size	Mean	SD	Minimum b	Maximum b
n20_20%	36	-0.10	0.92	-2.09	1.62
n20_30%	34	0.01	0.87	-1.69	1.80
n20_40%	32	-0.11	0.95	-2.13	1.85
n40_20%	72	0.12	0.98	-2.16	2.67
n40_30%	68	-0.01	1.01	-2.57	1.95
n40_40%	64	0.01	0.96	-2.33	2.17
n60_20%	108	-0.10	0.97	-2.56	2.27
n60_30%	102	-0.02	1.00	-2.62	2.50
n60_40%	96	0.00	1.09	-2.28	3.23

Table 4.2 Item pool descriptive stastitics of the MST 1-2 design with exposure control

Routing Proportion	Pool Size	Mean	SD	Minimum b	Maximum b
n20_20%	360	-0.10	0.87	-2.18	1.87
n20_30%	360	0.05	0.82	-2.27	2.25
n20_40%	360	0.05	0.75	-1.92	2.27
n40_20%	720	-0.02	0.93	-2.26	2.61
n40_30%	720	-0.01	0.94	-2.62	2.59
n40_40%	720	-0.01	0.76	-2.27	2.21
n60_20%	1080	0.00	0.93	-2.55	2.54
n60_30%	1080	-0.02	1.01	-2.93	2.58
n60_40%	1080	-0.02	0.99	-2.61	3.32

Table 4.3 Item pool descriptive stastitics of the MST 1-3 design without exposure control

Routing Proportion	Pool Size	Mean	SD	Minimum b	Maximum b
n20_20%	52	0.05	0.89	-2.00	2.24
n20_30%	48	-0.01	0.90	-1.99	2.14
n20_40%	44	-0.02	0.83	-2.04	1.76
n40_20%	104	-0.04	0.98	-2.59	2.43
n40_30%	96	-0.04	0.87	-2.17	2.11
n40_40%	88	-0.07	0.97	-2.79	2.52
n60_20%	156	-0.03	1.00	-2.52	2.33
n60_30%	144	-0.05	0.95	-2.80	2.19
n60_40%	132	0.02	0.99	-2.51	2.78

Table 4.4 Item pool descriptive stastitics of the MST 1-3 design with exposure control

Routing Proportion	Pool Size	Mean	SD	Minimum b	Maximum b
n20_20%	360	0.02	0.82	-2.25	2.25
n20_30%	360	0.03	0.76	-2.27	1.89
n20_40%	360	0.03	0.81	-2.51	2.26
n40_20%	720	-0.05	0.91	-2.27	2.55
n40_30%	720	0.03	0.86	-2.62	2.26
n40_40%	720	0.03	0.83	-2.44	2.94
n60_20%	1080	-0.06	0.98	-3.51	2.95
n60_30%	1080	0.01	0.93	-2.87	3.23
n60_40%	1080	0.01	0.87	-2.94	2.54

Table 4.5 Item pool descriptive stastitics of the MST 1-2-2 design without exposure control

Routing Proportion	Pool Size	Mean	SD	Minimum b	Maximum b
n20_20%	36	-0.06	0.79	-2.15	1.53
n20_30%	34	-0.05	0.72	-1.78	1.24
n20_40%	32	0.05	0.72	-1.43	1.57
n40_20%	72	-0.11	0.97	-2.86	2.35
n40_30%	68	-0.05	0.88	-2.59	2.00
n40_40%	64	0.03	0.93	-2.20	2.32
n60_20%	108	0.00	1.28	-2.97	2.90
n60_30%	102	0.03	1.32	-2.83	3.22
n60_40%	96	0.04	1.24	-2.51	2.67

Table 4.6 Item pool descriptive stastitics of the MST 1-2-2 design with exposure control

Routing Proportion	Pool Size	Mean	SD	Minimum b	Maximum b
n20_20%	360	0.00	0.75	-1.89	1.90
n20_30%	360	-0.05	0.67	-1.88	1.51
n20_40%	360	-0.05	0.64	-1.56	1.56
n40_20%	720	-0.04	0.91	-2.95	2.26
n40_30%	720	-0.02	0.79	-2.24	2.23
n40_40%	720	-0.02	0.81	-2.59	2.24
n60_20%	1080	-0.01	1.22	-2.92	3.26
n60_30%	1080	0.00	1.12	-2.60	2.97
n60_40%	1080	0.00	1.14	-2.96	3.31

Table 4.7 Item pool descriptive stastitics of the MST 1-2-3 design without exposure control

Routing Proportion	Pool Size	Mean	SD	Minimum b	Maximum b
n20_20%	44	-0.07	0.82	-1.66	1.66
n20_30%	42	0.01	0.77	-1.81	1.73
n20_40%	40	0.04	0.86	-2.16	2.01
n40_20%	88	-0.02	0.79	-1.99	2.10
n40_30%	84	-0.01	0.90	-2.42	2.59
n40_40%	80	-0.03	0.91	-2.47	2.10
n60_20%	132	-0.03	1.09	-3.34	2.96
n60_30%	126	-0.04	1.13	-2.76	2.65
n60_40%	120	-0.04	1.13	-2.65	2.57

Table 4.8 Item pool descriptive stastitics of the MST 1-2-3 design with exposure control

Routing Proportion	Pool Size	Mean	SD	Minimum b	Maximum b
n20_20%	360	-0.01	0.75	-1.88	1.87
n20_30%	360	0.03	0.65	-1.88	1.57
n20_40%	360	0.03	0.73	-2.26	2.26
n40_20%	720	-0.01	0.76	-2.95	2.55
n40_30%	720	-0.01	0.79	-2.60	2.54
n40_40%	720	-0.01	0.72	-2.55	2.13
n60_20%	1080	-0.02	0.93	-2.88	2.92
n60_30%	1080	-0.02	0.91	-2.58	3.29
n60_40%	1080	-0.02	0.93	-2.51	2.62

The absolute values of the differences between the mean, SD, minimum and maximum values of the descriptive statistics of the item pools under exposure control and non-exposure control conditions were calculated and compared. From the comparisons, it was concluded that no difference values between the means and standard deviations of the item pools under exposure control and non- exposure control conditions were larger than .3, and many were smaller than .01. The majority of the difference values between the minimum and maximum of the item pool parameters were smaller than .5 and some values were as small as .01. These results are expected since exposure control procedures enlarged the item pool size, but the item parameter characteristics were not supposed to change a lot. Based on the exposure control characteristics, a higher proportion of items were accumulated at the first stage as compared with no exposure control conditions. Thus the mean and variance of the item parameters after the implementation of exposure control could possibly vary. With more items being added to the pool, the maximum and minimum values of the item parameters could also change. After exposure control was implemented, the item pool sizes under certain test length became the same. This is because the characteristics of the inverse proportional method made the item pool a matrix of test form number * test length. For example, for the MST 1-2 design under 20-item test, it is 18 (e.g., for first stage and second stage) *20. So even the routing test proportion was different under certain test length, after the exposure control, they all shared the same item pool size. Comparatively speaking, with no exposure control implementation, the item pool sizes for different routing test proportions under the same test length was different.

A ratio of the item pool size between exposure control and no exposure control conditions were calculated and a two-way ANOVA was conducted to see the association between item pool size, MST designs and routing test proportions. Since test length has no

variance on item pool size change, it was excluded from the analysis. The interaction term between MST design and routing test proportions was used as the error term in the model (Winer et al., 1991). Table 4.9 showed the values of the variables and the ratio in the two-way ANOVA. Table 4.10 showed the results of the two-way ANOVA analysis on item pool size comparisons.

It was discovered that the main effects of MST design, F (3, 30) = 2975.90, p=.00, and routing test proportions, F (2, 30) = 551.93, p=.00 were both statistically significant.
The null hypotheses were rejected and the ratio of the item pool size significantly differed from each other across the different MST designs and routing test proportions. Generally speaking, for MST 1-2 design, the item pool size under exposure control was about 10-11 times larger than those under no exposure control; for MST 1-3 design, the item pool size under exposure control was about 7-8 times larger than those under no exposure control; for MST 1-2-2 design, the item pool size under exposure control was about 10-11 times larger than those under no exposure control; for MST 1-2-3 design, the item pool size under exposure control was about 8-9 times larger than those under no exposure control. Variation occurred when the routing test proportion was different.

Table 4.9 Ratio of the item pool size between exposure control and non-exposure control

MST Design	Test Length	Routing Length	Ratio
MST1-2	20	0.2	10
MST1-2	20	0.3	11
MST1-2	20	0.4	11
MST1-2	40	0.2	10
MST1-2	40	0.3	11
MST1-2	40	0.4	11
MST1-2	60	0.2	10
MST1-2	60	0.3	11
MST1-2	60	0.4	11
MST1-3	20	0.2	7
MST1-3	20	0.3	8
MST1-3	20	0.4	8
MST1-3	40	0.2	7
MST1-3	40	0.3	8
MST1-3	40	0.4	8
MST1-3	60	0.2	7
MST1-3	60	0.3	8
MST1-3	60	0.4	8
MST1-2-2	20	0.2	10
MST1-2-2	20	0.3	11
MST1-2-2	20	0.4	11
MST1-2-2	40	0.2	10
MST1-2-2	40	0.3	11
MST1-2-2	40	0.4	11
MST1-2-2	60	0.2	10
MST1-2-2	60	0.3	11
MST1-2-2	60	0.4	11
MST1-2-3	20	0.2	8
MST1-2-3	20	0.3	9
MST1-2-3	20	0.4	9
MST1-2-3	40	0.2	8
MST1-2-3	40	0.3	9
MST1-2-3	40	0.4	9
MST1-2-3	60	0.2	8
MST1-2-3	60	0.3	9
MST1-2-3	60	0.4	9

Note: The ratio is a rounded up value calculated by dividing the item pool size under the exposure control condition by the no exposure control conditions.

Table 4.10 Two-Way ANOVA results on item pool size comparisons

Source	Sum of Squares	df	Mean Square	F	P
Corrected Model	71.49[a]	5	14.30	2006.31	.00
Intercept	3137.94	1	3137.94	440349.90	.00
MST design	63.62	3	21.21	2975.90	.00*
Routing length	7.87	2	3.93	551.93	.00*
Error	.21	30	.01		
Total	3209.64	36			
Corrected Total	71.70	35			

a. *R* Squared = .997 (Adjusted *R* Squared = .997); *p**<.001

Table 4.11 to Table 4.18 showed the evaluation results by the overall sample for within all the four MST designs under the conditions of without exposure control and with exposure control. As in Figure 4.26, n20_20% means the test length is 20 and routing test proportion is 20%. The results indicated that the correlations between true latent ability and estimated ability were all very high for all test configurations in all MST designs. The overall bias and RMSE were all quite small. Under the same routing test length proportion, with the increase of test length, the RMSE and standard error decreased. The overall test information, reliability and classification accuracy all increased. With the increase of test length, the classification accuracy as indicated by the three cutoff scores became more accurate. The reliability for all test designs and configurations all exceeded .90. These results applied to both conditions for without exposure control and with exposure control. When no exposure control was implemented, the item overlap rates for all the conditions across all the MST designs were quite big. Under exposure control condition, the item overlap rates all became quite small, and the same was true with item exposure rate conditional on bin. Under the no exposure control condition, no item had exposure rate larger than .20 within each bin. Under the exposure control condition, no item had exposure rate larger than .20 and smaller than .02.

Table 4.11 The performance of the MST 1-2 optimal item pool without exposure control

Proportion	Corr	Bias	RMSE	Infor	SE	R	Class Median	Class Competent	Class Scholar	Overlap rate	Exposure rate
n20_20%	1.00	0.01	0.35	9.72	0.32	0.92	0.90	0.93	0.97	0.61	0.52
n40_20%	1.00	0.00	0.24	19.44	0.23	0.95	0.92	0.95	0.98	0.60	0.53
n60_20%	1.00	0.00	0.19	30.01	0.18	0.97	0.94	0.96	0.98	0.60	0.53
n20_30%	1.00	0.00	0.35	9.88	0.32	0.92	0.90	0.93	0.97	0.65	0.52
n40_30%	1.00	0.00	0.24	19.41	0.23	0.95	0.93	0.95	0.98	0.65	0.52
n60_30%	1.00	0.00	0.19	29.71	0.18	0.97	0.94	0.96	0.98	0.65	0.53
n20_40%	1.00	0.01	0.35	9.69	0.32	0.92	0.91	0.93	0.96	0.70	0.52
n40_40%	1.00	0.00	0.24	19.77	0.22	0.95	0.93	0.95	0.98	0.70	0.52
n60_40%	1.00	0.00	0.19	29.24	0.18	0.97	0.94	0.96	0.98	0.70	0.53

Note.Corr=Correlation; Infor=Information; SE=standard error of measurement; R=marginal reliability; Class Median/Competent/Scholar=Classification accuracy based on median/minimum competence/scholarship cutoff scores

Table 4.12 The performance of the MST 1-2 optimal item pool with exposure control

Proportion	Corr	Bias	RMSE	Infor	SE	R	Class Median	Class Competent	Class Scholar	Overlap rate	Exposure rate
n20_20%	1.00	0.00	0.36	16.48	0.25	0.94	0.90	0.93	0.97	0.06	0.04
n40_20%	1.00	0.00	0.26	30.48	0.18	0.96	0.92	0.94	0.98	0.06	0.04
n60_20%	1.00	0.00	0.22	43.03	0.15	0.96	0.94	0.96	0.98	0.06	0.04
n20_30%	1.00	0.00	0.37	14.00	0.27	0.93	0.91	0.92	0.97	0.06	0.04
n40_30%	1.00	0.00	0.27	27.63	0.19	0.95	0.93	0.95	0.98	0.06	0.04
n60_30%	1.00	0.00	0.23	40.89	0.16	0.96	0.94	0.95	0.98	0.06	0.03
n20_40%	1.00	-0.01	0.38	13.03	0.28	0.93	0.90	0.93	0.97	0.06	0.04
n40_40%	1.00	0.00	0.27	26.82	0.19	0.95	0.93	0.94	0.98	0.06	0.03
n60_40%	1.00	0.00	0.22	39.51	0.16	0.97	0.94	0.96	0.98	0.06	0.04

Note.Corr=Correlation; Infor=Information; SE=standard error of measurement; R=marginal reliability; Class Median/Competent/Scholar=Classification accuracy based on median/minimum competence/scholarship cutoff scores

Table 4.13 The performance of the MST 1-3 optimal item pool without exposure control

Proportion	Corr	Bias	RMSE	Infor	SE	R	Class Median	Class Competent	Class Scholar	Overlap rate	Exposure rate
n20_20%	1.00	-0.01	0.34	9.84	0.32	0.92	0.90	0.93	0.97	0.48	0.38
n40_20%	1.00	0.00	0.23	20.36	0.22	0.96	0.94	0.95	0.98	0.47	0.41
n60_20%	1.00	0.00	0.18	31.86	0.18	0.97	0.94	0.96	0.98	0.47	0.39
n20_30%	1.00	0.00	0.34	9.80	0.32	0.92	0.90	0.93	0.98	0.56	0.35
n40_30%	1.00	0.00	0.23	20.84	0.22	0.95	0.93	0.95	0.98	0.53	0.37
n60_30%	1.00	0.01	0.18	31.81	0.18	0.97	0.95	0.96	0.99	0.53	0.38
n20_40%	1.00	0.00	0.34	10.37	0.31	0.92	0.91	0.94	0.97	0.60	0.36
n40_40%	1.00	0.00	0.23	20.72	0.22	0.96	0.93	0.95	0.98	0.60	0.36
n60_40%	1.00	0.00	0.19	31.14	0.18	0.97	0.94	0.96	0.98	0.60	0.39

Note.Corr=Correlation; Infor=Information; SE=standard error of measurement; R=marginal reliability; Class Median/Competent/Scholar=Classification accuracy based on median/minimum competence/scholarship cutoff scores

Table 4.14 The performance of the MST 1-3 optimal item pool with exposure control

Proportion	Corr	Bias	RMSE	Infor	SE	R	Class Median	Class Competent	Class Scholar	Overlap rate	Exposure rate
n20_20%	1.00	0.00	0.34	9.80	0.32	0.90	0.90	0.93	0.97	0.06	0.06
n40_20%	1.00	0.00	0.23	20.55	0.22	0.95	0.93	0.95	0.98	0.06	0.06
n60_20%	1.00	0.00	0.19	30.98	0.18	0.97	0.95	0.96	0.99	0.06	0.05
n20_30%	1.00	0.00	0.34	9.88	0.32	0.90	0.90	0.94	0.98	0.06	0.06
n40_30%	1.00	0.00	0.23	20.69	0.22	0.95	0.93	0.95	0.98	0.06	0.06
n60_30%	1.00	0.00	0.19	31.26	0.18	0.97	0.95	0.96	0.98	0.06	0.06
n20_40%	1.00	0.00	0.34	10.08	0.32	0.90	0.91	0.93	0.97	0.06	0.06
n40_40%	1.00	0.00	0.23	20.55	0.22	0.95	0.93	0.95	0.98	0.06	0.05
n60_40%	1.00	0.00	0.19	31.30	0.18	0.97	0.94	0.96	0.99	0.06	0.06

Note.Corr=Correlation; Infor=Information; SE=standard error of measurement; R=marginal reliability; Class Median/Competent/Scholar=Classification accuracy based on median/minimum competence/scholarship cutoff scores

Table 4.15 The performance of the MST 1-2-2 optimal item pool without exposure control

Proportion	Corr	Bias	RMSE	Infor	SE	R	Class Median	Class Competent	Class Scholar	Overlap rate	Exposure rate
n20_20%	1.00	0.01	0.36	9.68	0.32	0.92	0.90	0.93	0.97	0.61	0.54
n40_20%	1.00	0.00	0.24	19.04	0.23	0.95	0.93	0.95	0.98	0.60	0.53
n60_20%	1.00	0.00	0.19	27.31	0.19	0.96	0.93	0.96	0.98	0.60	0.54
n20_30%	1.00	0.00	0.37	9.70	0.32	0.92	0.91	0.93	0.97	0.65	0.52
n40_30%	1.00	0.00	0.24	19.05	0.23	0.95	0.92	0.95	0.98	0.65	0.53
n60_30%	1.00	0.00	0.20	26.67	0.19	0.96	0.93	0.96	0.99	0.65	0.53
n20_40%	1.00	0.00	0.39	9.50	0.32	0.93	0.91	0.93	0.97	0.70	0.53
n40_40%	1.00	0.00	0.25	18.45	0.23	0.95	0.93	0.95	0.98	0.70	0.53
n60_40%	1.00	0.00	0.20	26.56	0.19	0.96	0.94	0.96	0.98	0.70	0.54

Note.Corr=Correlation; Infor=Information; SE=standard error of measurement; R=marginal reliability; Class Median/Competent/Scholar=Classification accuracy based on median/minimum competence/scholarship cutoff scores

Table 4.16 The performance of the MST 1-2-2 optimal item pool with exposure control

Proportion	Corr	Bias	RMSE	Infor	SE	R	Class Median	Class Competent	Class Scholar	Overlap rate	Exposure rate
n20_20%	1.00	0.00	0.36	9.70	0.32	0.92	0.89	0.93	0.97	0.06	0.05
n40_20%	1.00	0.00	0.24	19.00	0.23	0.95	0.93	0.95	0.98	0.06	0.05
n60_20%	1.00	0.00	0.20	27.14	0.19	0.97	0.94	0.96	0.99	0.06	0.06
n20_30%	1.00	0.00	0.37	9.65	0.32	0.93	0.91	0.93	0.96	0.06	0.05
n40_30%	1.00	0.00	0.25	19.06	0.23	0.95	0.93	0.95	0.98	0.06	0.05
n60_30%	1.00	0.00	0.20	27.10	0.19	0.96	0.94	0.96	0.98	0.06	0.05
n20_40%	1.00	0.01	0.39	9.49	0.32	0.93	0.90	0.92	0.96	0.06	0.05
n40_40%	1.00	0.00	0.25	18.56	0.23	0.95	0.93	0.94	0.98	0.06	0.05
n60_40%	1.00	0.00	0.20	26.22	0.20	0.96	0.94	0.95	0.99	0.06	0.05

Note.Corr=Correlation; Infor=Information; SE=standard error of measurement; R=marginal reliability; Class Median/Competent/Scholar=Classification accuracy based on median/minimum competence/scholarship cutoff scores

Table 4.17 The performance of the MST 1-2-3 optimal item pool without exposure control

Proportion	Corr	Bias	RMSE	Infor	SE	R	Class Median	Class Competent	Class Scholar	Overlap rate	Exposure rate
n20_20%	1.00	-0.01	0.33	10.75	0.31	0.93	0.90	0.93	0.97	0.54	0.37
n40_20%	1.00	0.00	0.23	21.78	0.21	0.96	0.94	0.95	0.98	0.53	0.39
n60_20%	1.00	0.00	0.19	31.16	0.18	0.97	0.95	0.96	0.98	0.53	0.38
n20_30%	1.00	0.00	0.34	10.68	0.31	0.93	0.91	0.93	0.97	0.59	0.38
n40_30%	1.00	0.00	0.23	21.29	0.22	0.96	0.94	0.95	0.98	0.58	0.37
n60_30%	1.00	0.00	0.19	30.24	0.18	0.97	0.94	0.96	0.98	0.59	0.38
n20_40%	1.00	0.01	0.34	10.36	0.31	0.92	0.90	0.93	0.97	0.64	0.39
n40_40%	1.00	0.00	0.23	20.90	0.22	0.96	0.94	0.95	0.98	0.63	0.38
n60_40%	1.00	0.00	0.19	29.64	0.18	0.97	0.94	0.96	0.98	0.63	0.39

Note.Corr=Correlation; Infor=Information; SE=standard error of measurement; R=marginal reliability; Class Median/Competent/Scholar=Classification accuracy based on median/minimum competence/scholarship cutoff scores

Table 4.18 The performance of the MST 1-2-3 optimal item pool with exposure control

Proportion	Corr	Bias	RMSE	Infor	SE	R	Class Median	Class Competent	Class Scholar	Overlap rate	Exposure rate
n20_20%	1.00	0.00	0.33	10.80	0.30	0.92	0.90	0.93	0.97	0.06	0.05
n40_20%	1.00	0.00	0.23	21.70	0.21	0.95	0.93	0.95	0.98	0.06	0.05
n60_20%	1.00	0.00	0.19	31.18	0.18	0.97	0.94	0.96	0.99	0.06	0.05
n20_30%	1.00	0.00	0.34	10.75	0.31	0.93	0.91	0.94	0.97	0.06	0.05
n40_30%	1.00	0.00	0.23	21.20	0.22	0.95	0.93	0.95	0.98	0.06	0.05
n60_30%	1.00	0.00	0.19	30.48	0.18	0.97	0.95	0.96	0.98	0.06	0.05
n20_40%	1.00	0.00	0.34	10.35	0.31	0.92	0.91	0.92	0.97	0.06	0.06
n40_40%	1.00	0.00	0.23	21.08	0.22	0.96	0.93	0.95	0.98	0.06	0.05
n60_40%	1.00	0.00	0.19	29.58	0.18	0.97	0.94	0.96	0.99	0.06	0.06

Note.Corr=Correlation; Infor=Information; SE=standard error of measurement; R=marginal reliability; Class Median/Competent/Scholar=Classification accuracy based on median/minimum competence/scholarship cutoff scores

4.2.2 Results across the different MST designs

The results for comparisons of classification accuracy across the different MST designs were presented in Figure 4.27 to Figure 4.32 for both without exposure control and with exposure control conditions. The meanings of the labels have the similar meanings with the labels in Figure 4.26. For example, MST12.n40.20% means the condition of MST 1-2 design with a test length of 40 and routing test proportion of 20%. M means module. Based on the 95% correct classification rates, it was concluded that the MST 1-2-3 design was the best in terms of the classification accuracy for the median cutoff score at a test length of 20 for all routing test proportions, and the MST 1-3 worked the best for the minimum competence and scholarship cutoff scores under the same conditions. At the test length of 40, except for the routing proportion of 30%, the MST 1-2-3 design also worked the best for the median cutoff score. There is some variation for the minimum competence and scholarship cutoff scores. For the minimum competence cutoffs core, the MST 1-3 design was the best when the routing proportion is 20%, and the MST 1-2-3 design worked the best when the routing proportions were 30% and 40%. For the scholarship cutoff score, the MST 1-2-2 design worked the best when the routing proportions were 20% and 30%, and the MST 1-2 design had the highest classification accuracy when the routing proportion was 40%. At the test length of 60, the MST 1-2-3 design worked the best for the median cutoff score when the routing proportions were 20% and 30%, and the MST 1-3 design had higher classification accuracy when the routing test proportion was 40%. Variation continued to occur for the minimum competence and scholarship cutoff scores. For the minimum competence cutoff score, the MST 1-2-3 design had the highest classification accuracy when the routing test proportion was 20%, and the best design was switched to the MST 1-2-2 design when the routing test proportion was 30%, and the MST 1-2 design when the

routing proportion was 40%. At the scholarship cutoff score, the MST 1-2-3 had the highest classification accuracy when the routing test proportion was 20% and 30%, and the MST 1-3 design had the highest classification accuracy when the routing test was increased to 40%.

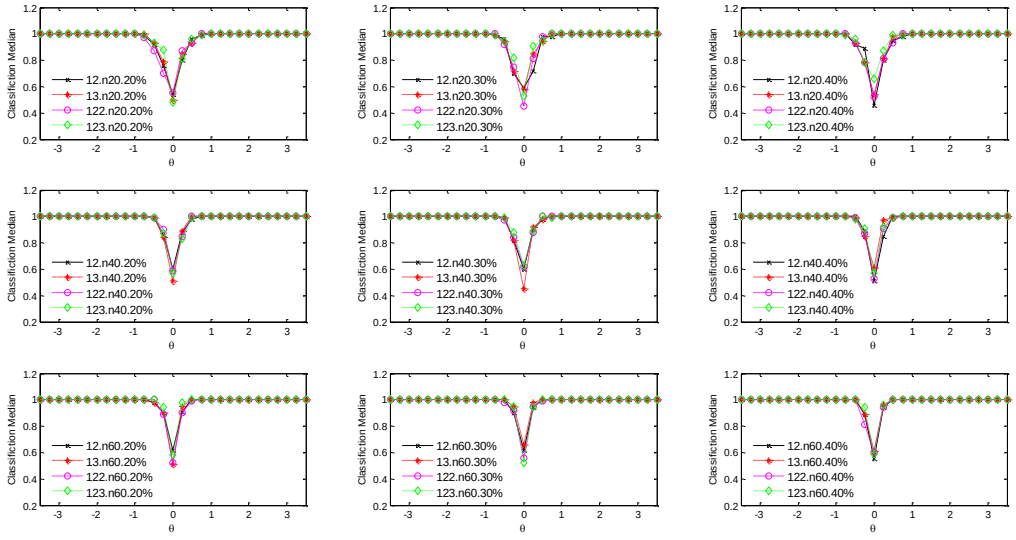

Figure 4.27 Classification accuracy for median cutoff scores across MST designs without exposure control

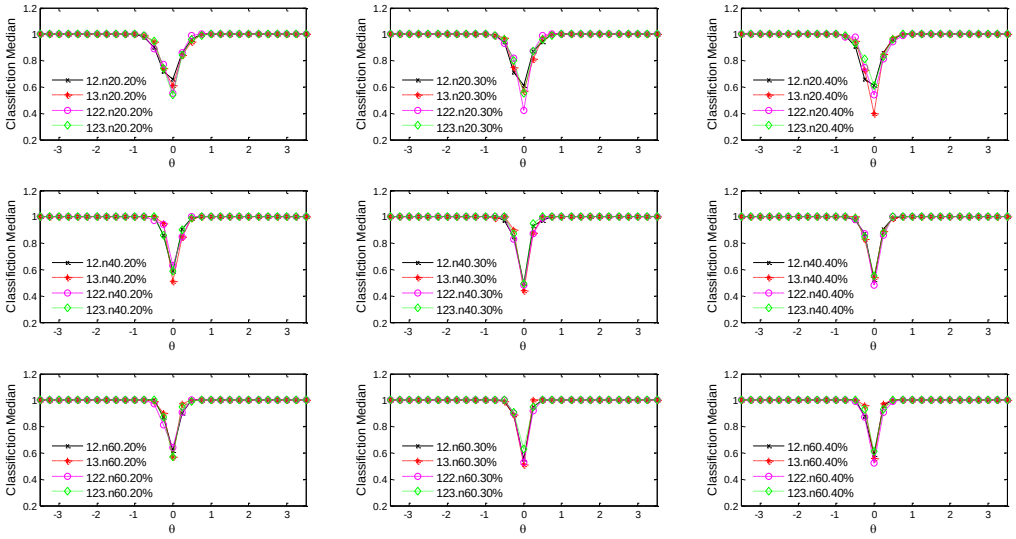

Figure 4.28 Classification accuracy for median cutoff scores across MST designs with exposure control

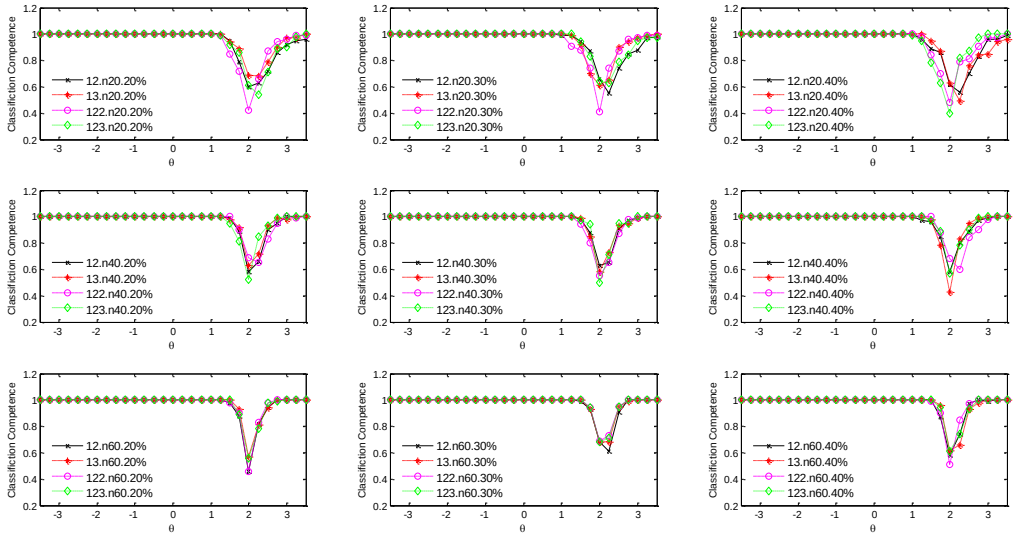

Figure 4.29 Classification accuracy for minimum competence cutoff scores across MST designs without exposure control

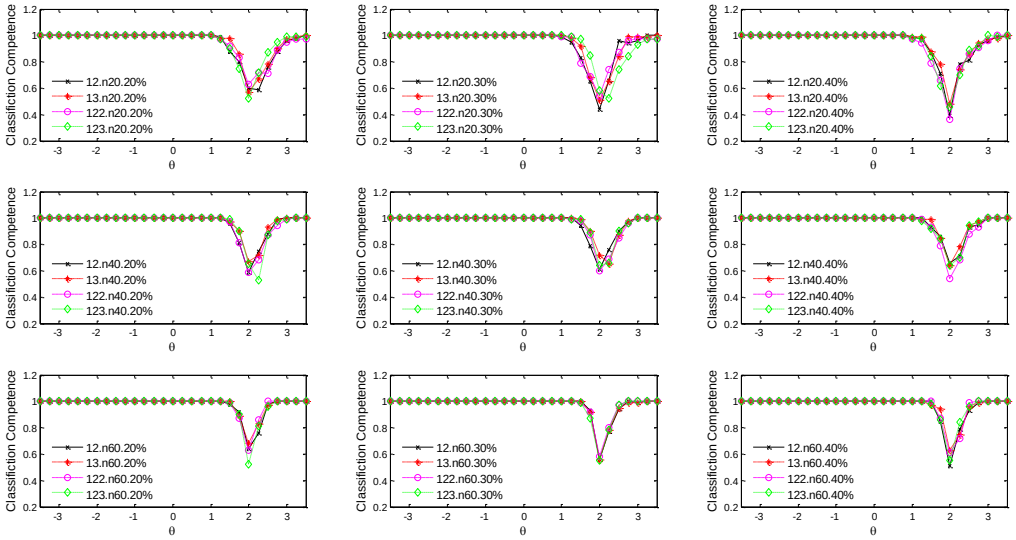

Figure 4.30 Classification accuracy for minimum competence cutoff scores across MST designs with exposure control

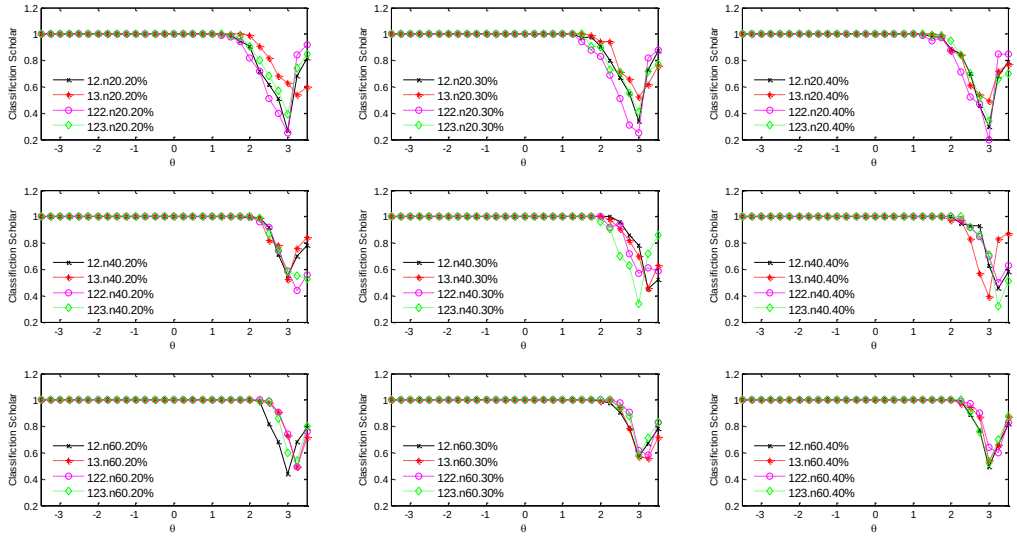

Figure 4.31 Classification accuracy for scholarship cutoff scores across MST designs without exposure control

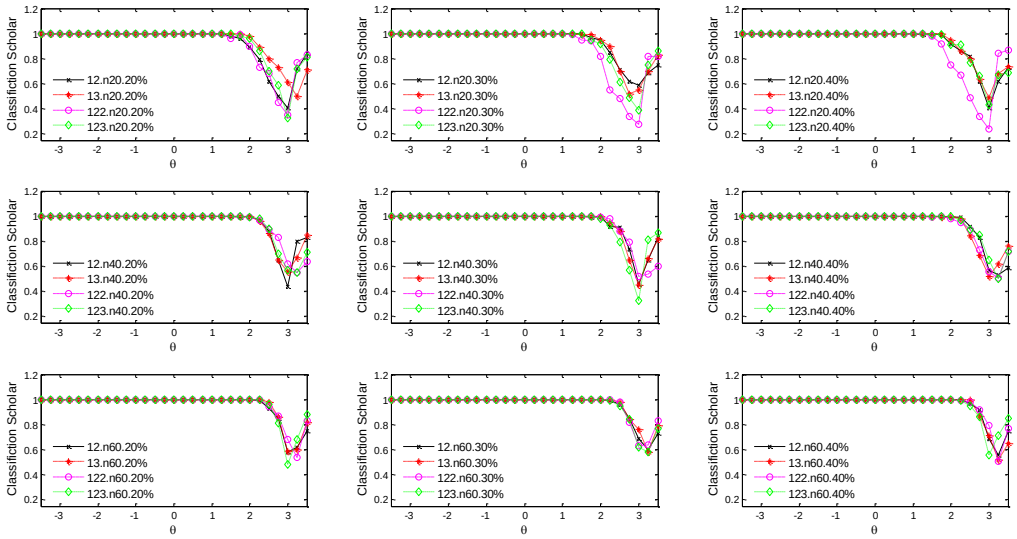

Figure 4.32 Classification accuracy for scholarship cutoff scores across MST designs with exposure control

Figure 4.33 to Figure 4.40 display the comparison results across the different MST designs for conditional bias, RMSE, SE and item overlap rate under no exposure control and

exposure control conditions. The differences across different MST designs, different test lengths and routing test proportions under both exposure control and no exposure control conditions were compared. When the test length is 20, there were some differences at the two extremes across different MST designs. For example, the MST 1-2-2 design had higher bias, RMSE and SE than the rest of the designs. Not much difference was discovered for the rest of the test lengths for all the other conditions. Since more items were accumulated in the middle range of the ability scale, the conditional item overlap rate was the lowest in the middle and highest at the two extremes across all MST designs. At the two extremes, under no exposure control condition, since there was only one test form for all the examinees, it was likely that they all shared the same set of items. Thus the conditional item overlap rate was almost 1. For the MST 1-2-3 design, since there were more stages compared with two-stage design, and more modules compared with the MST 1-2-2 design, the item overlap rate was comparatively lower than the rest of the designs. But not many differences were discovered for all the designs under the exposure control condition except that the MST 1-3 design had slightly higher conditional overlap rate than the other three designs. In addition, the conditional item overlap rate was quite high for without exposure control conditions, but fell within the ideal range of .02 to .20 after exposure control was implemented. Slightly better measurement accuracy was discovered for some conditions under exposure control than no exposure control conditions.

As is shown from the following figures, for within the different MST designs, the conditional bias, RMSE and SE were higher at the lower end and higher end of the θ scale, but close to zero in the middle range of the θ scale. The conditional bias decreased with the increase of test length. The bias was the lowest when the test length is 60 and the highest when the test length is 20, especially at the two extremes where negative bias was found at the lower end and

positive bias was found at the higher end. Similar conditions applied to the conditional RMSE and SE. The values of RMSE increased with the decrease of test length and became the largest when the test length is 20. The conditional SE shared the same characteristics. These results applied to all the conditions for both without exposure control and with exposure control.

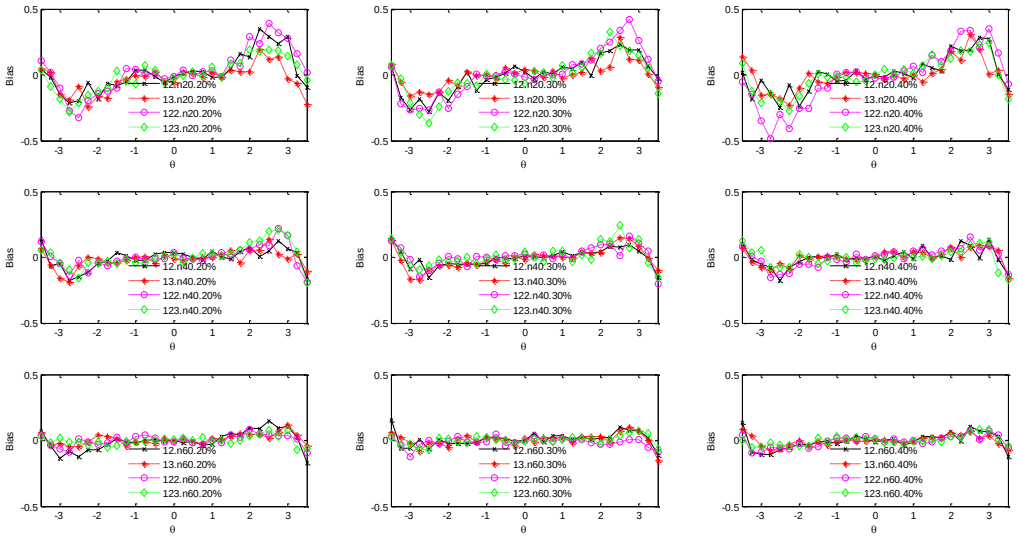

Figure 4.33 Conditional bias across MST designs without exposure control

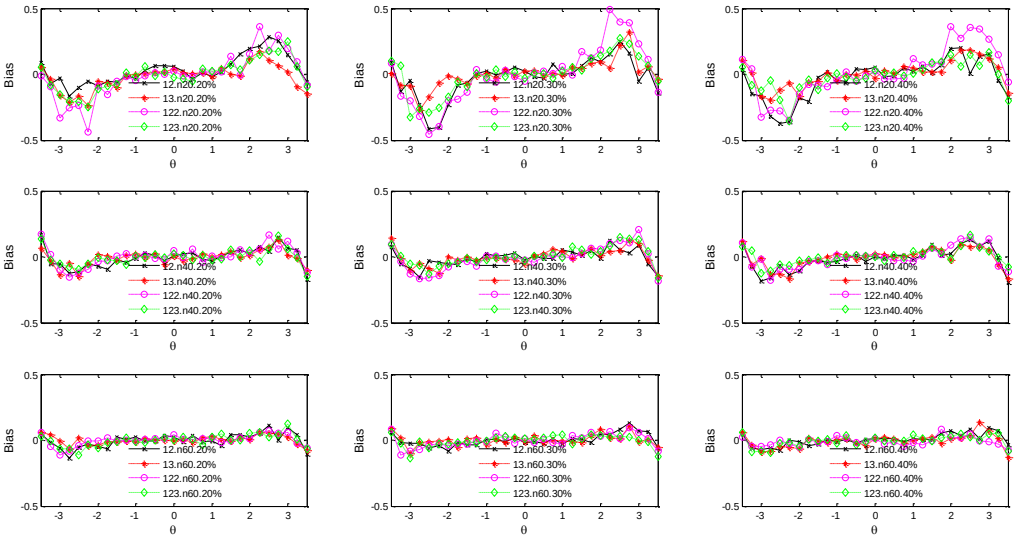

Figure 4.34 Conditional bias across MST designs with exposure control

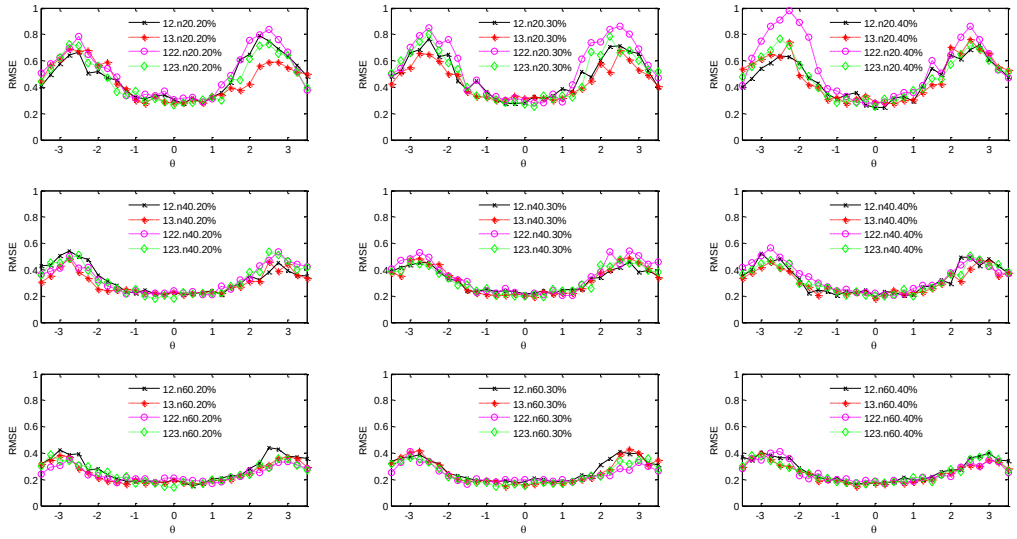

Figure 4.35 Conditional RMSE across MST designs without exposure control

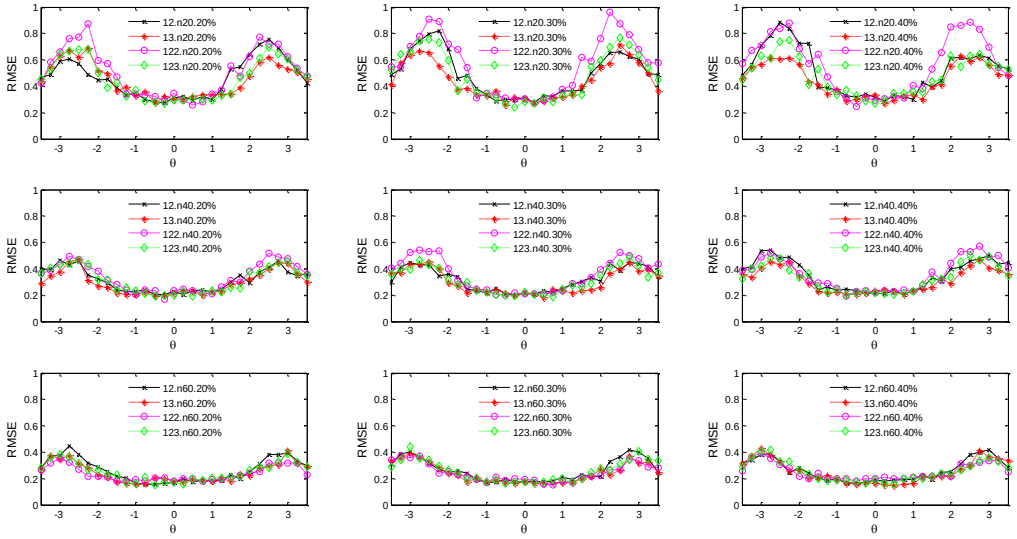

Figure 4.36 Conditional RMSE across MST designs with exposure control

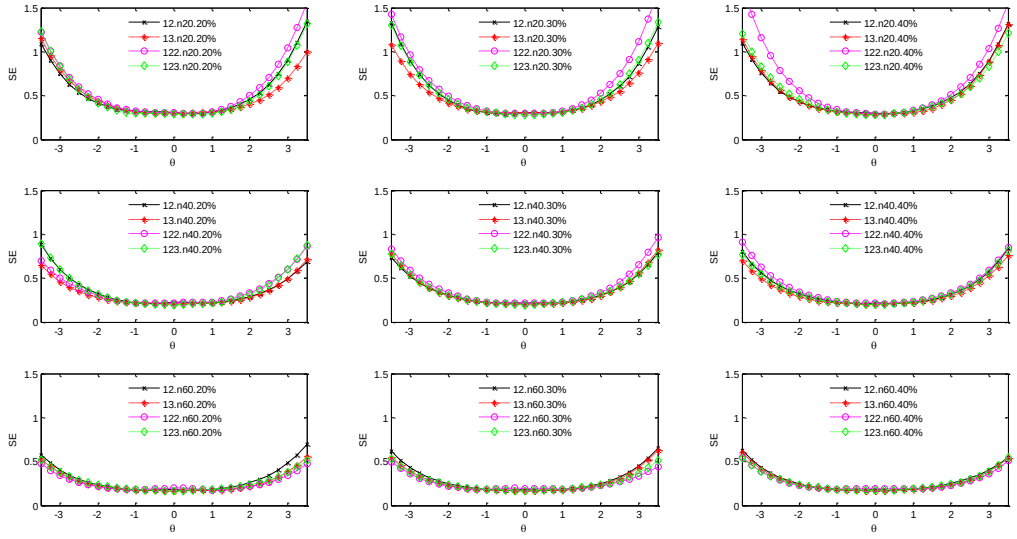

Figure 4.37 Conditional standard error across MST designs without exposure control

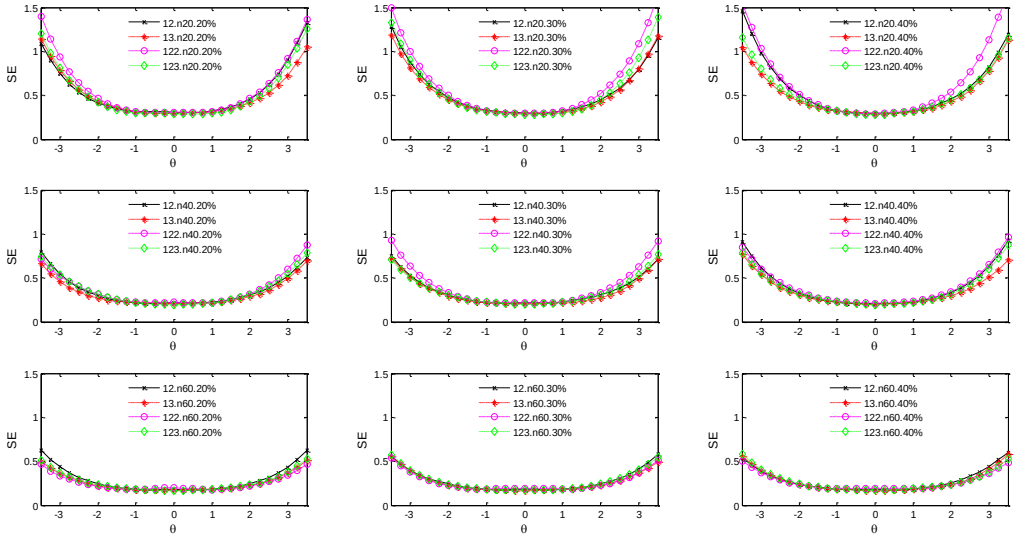

Figure 4.38 Conditional standard error across MST designs with exposure control

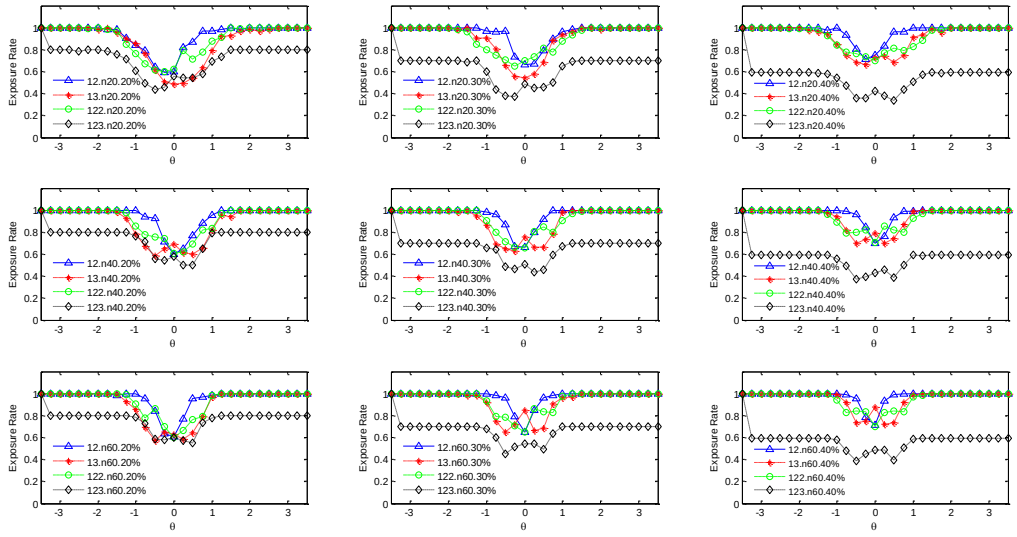

Figure 4.39 Conditional item overlap rate across MST designs without exposure control

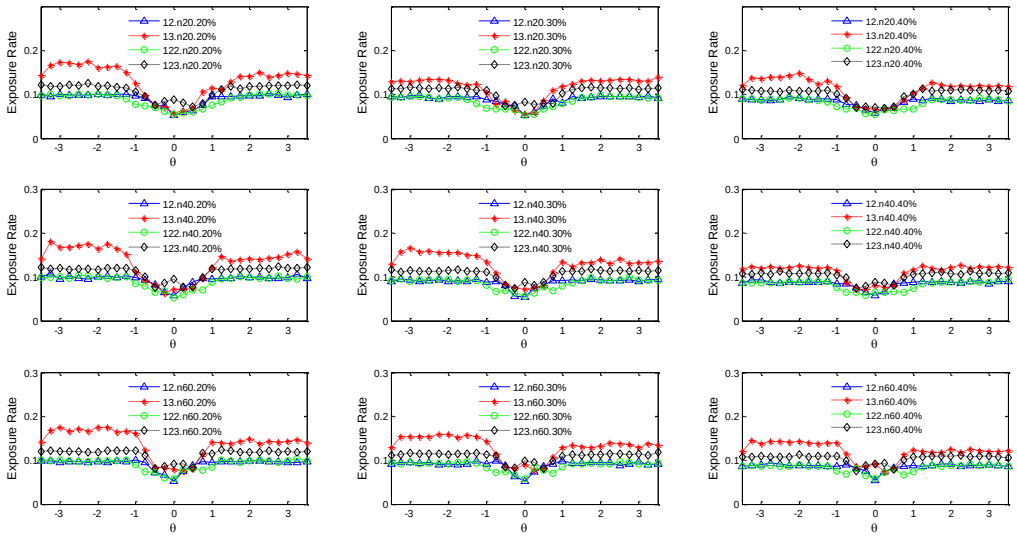

Figure 4.40 Conditional item overlap rates across MST designs with exposure control

4.3 Results from the application of the *p*-optimality method in an operational MST

The results from the simulated pool (*S*-Pool) and real pool (*R*-Pool) are summarized below. It includes the results from no exposure control conditions, and under exposure control conditions.

Figure 4.41 and Figure 4.42 showed the results of the number of items within each bin for all the stages in the MST 1-2-2 design for *S*-Pool and *R*-Pool under no exposure and exposure control conditions. Basically the item frequency distributions were symmetric. Since each module included 25 items and the modules for the first stage and second stage were mainly used for correctly classifying the examinees into the next stage, items were centered around the cutoff scores at those stages, and the information was not as spread as that at the third stage.

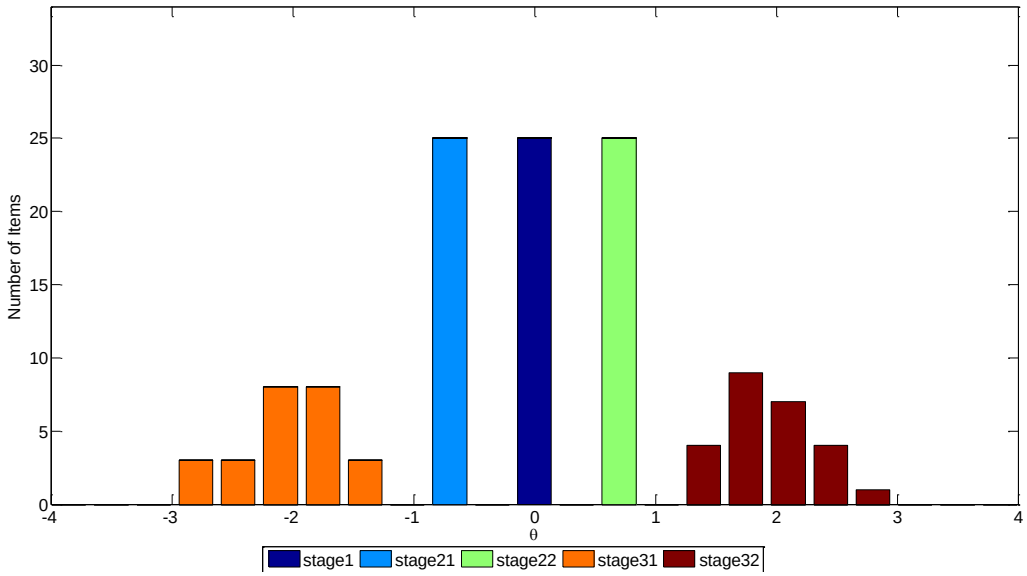

Figure 4.41 The frequency of items in the *R*-Pool and *S*-Pool without exposure control

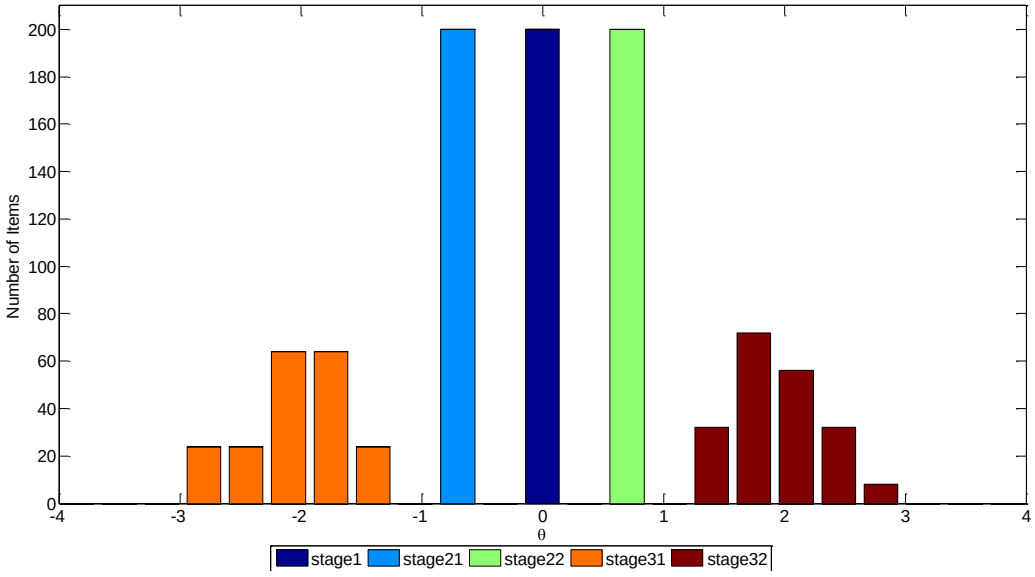

Figure 4.42 The frequency of items in the *R*-Pool and *S*-Pool with exposure control

Figure 4.43 shows the overlap of the item difficulty range across modules at Stage 3 for both *R*-Pool and *S*-Pool at the item pool design stage. Under the exposure control condition, the number of items within each bin for each module was multiplied by the number of test forms (e.g., by 8) that was required. Since the boundaries of the bins were the same with the non-exposure control condition, the item overlaps across modules under the exposure control condition were not displayed. Unlike the 1-2-2 design in the simulation part discussed above, Figure 4.43 showed no overlapping of items across the modules at Stage 3. This is because the operational test contains 25 items in each module. After the examinees took 50 items for the first stage and second stage, enough information were provided at the middle range of the θ scale and examinees were classified more accurately into the third stage modules.

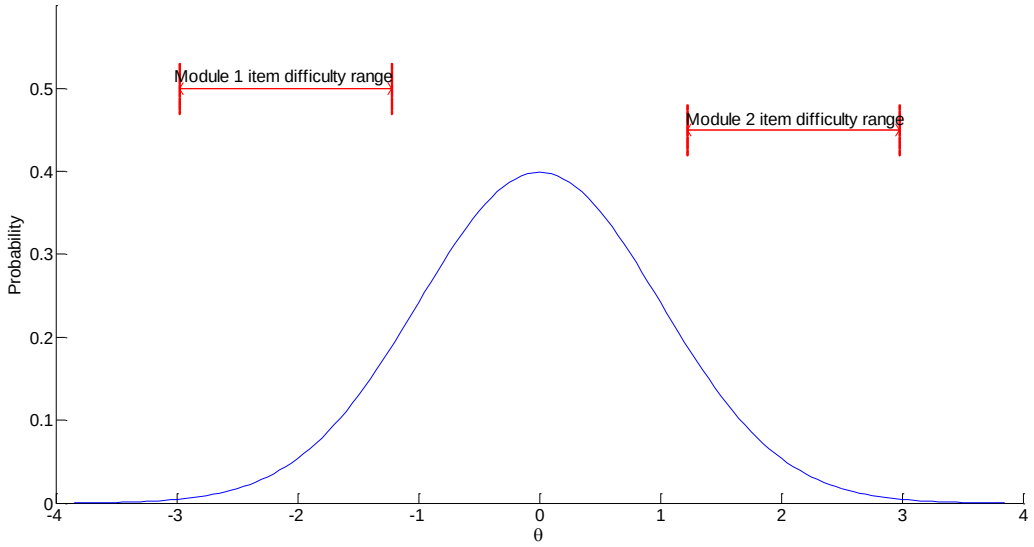

Figure 4.43 Item overlap across modules at Stage 3 for the *R*-Pool and *S*-Pool at the item pool design stage

Figure 4.44 shows the overlap of the item difficulty range across modules at Stage 3 for the *R*-Pool at the simulated item pool stage.

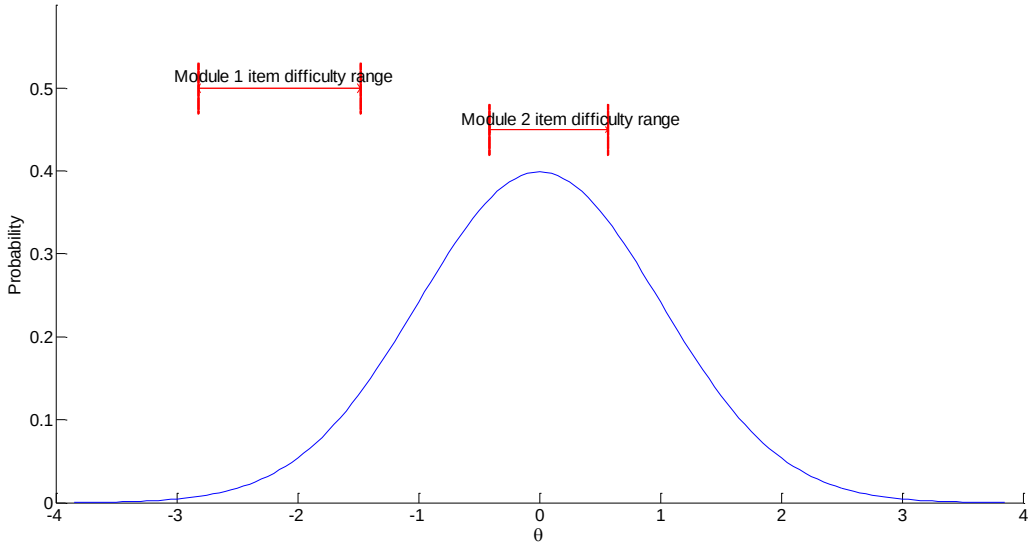

Figure 4.44 Item overlap across modules at Stage 3 for the *R*-Pool at the simulated item pool stage without exposure control

The results showed that most of the item parameters in the real pool covered the examinees' abilities below 1. Figure 4.45 showed the item overlap across modules at Stage 3 for the *S*-Pool at the simulated item pool stage without exposure control.

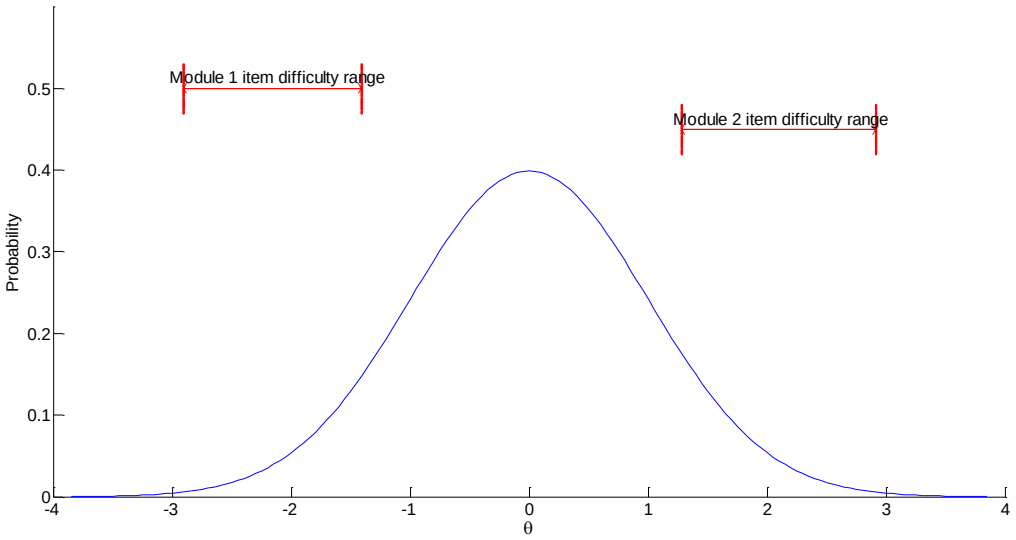

Figure 4.45 Item overlap across modules at Stage 3 for the *S*-Pool at the simulated item pool stage without exposure control

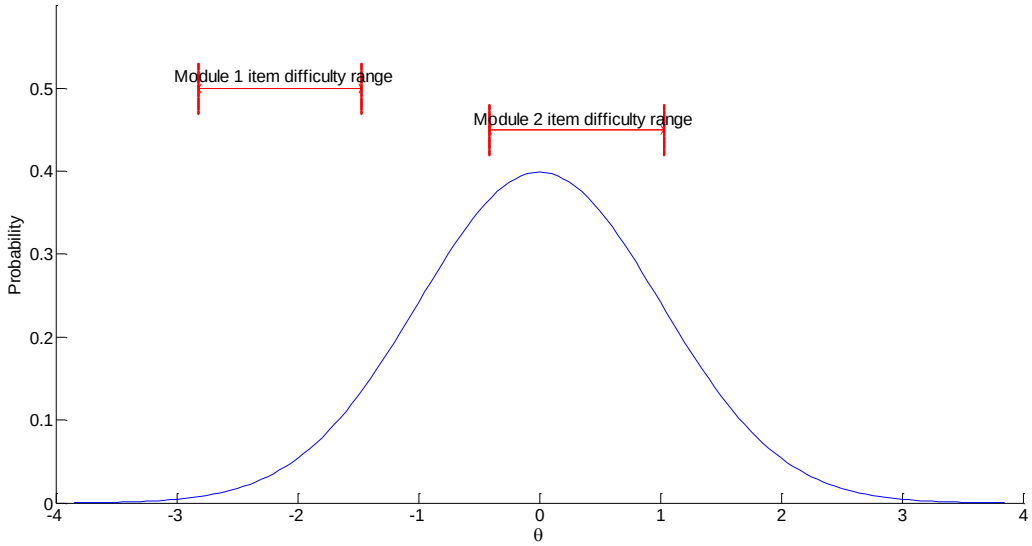

Figure 4.46 Item overlap across modules at Stage 3 for the *R*-Pool at the simulated item pool stage with exposure control

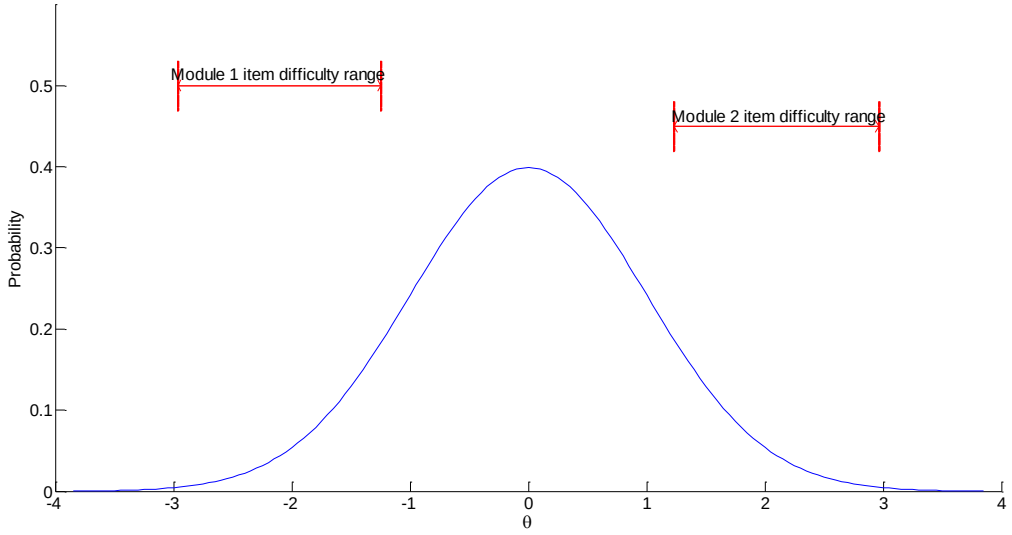

Figure 4.47 Item overlap across modules at Stage 3 for the *S*-Pool at the simulated item pool stage with exposure control

Figure 4.46 and figure 4.47 show the results of item overlap under exposure control conditions. Comparatively speaking, the items in the *S*-Pool covered a wider range of ability levels and were more aligned with the item parameter requirement at the item pool design stage.

Figure 4.48 shows the information function at the item pool design stage under no exposure control condition. As noted in Figure 4.48, the information function distribution was bimodal, which was determined by the features and requirements of the MST 1-2-2 design. In the middle of the information function, the line was jagged. Because 100 replications were conducted at the item pool design stage in order to reduce sampling error, it was possible that examinees were routed to different modules in different iterations. This possibility was more likely to occur with examinees in the middle range of the θ scale. After taking the average of the 100 replications, the middle range of the information function curve became jagged. The information function exceeded the required test information.

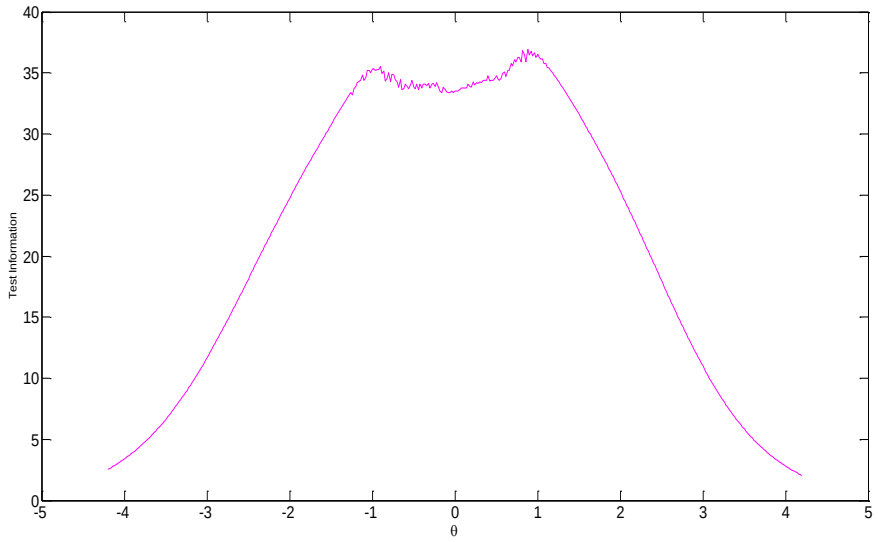

Figure 4.48 Test information function at the item pool design stage

Figure 4.49 shows the module information curves for the MST 1-2-2 design for both types of item pools under the no exposure control situation. The label M1 means module at Stage 1, M21 means the easy module at Stage 2, M22 means the hard module at Stage 2, and M31 means the easy module at Stage 3, and M32 means the hard module at Stage 3. It is observed from the figures that the modules at the same stages for the *S*-Pool were all symmetric, which means the modules built using the *p*-optimality method were parallel with each other. The modules for the *R*-Pool covered a narrower range in the ability scale than those for the *S*-Pool. The test information functions and module information functions in the following figures all indicated that the item pool parameters selected using the *p*-optimality method were appropriate for the MST design.

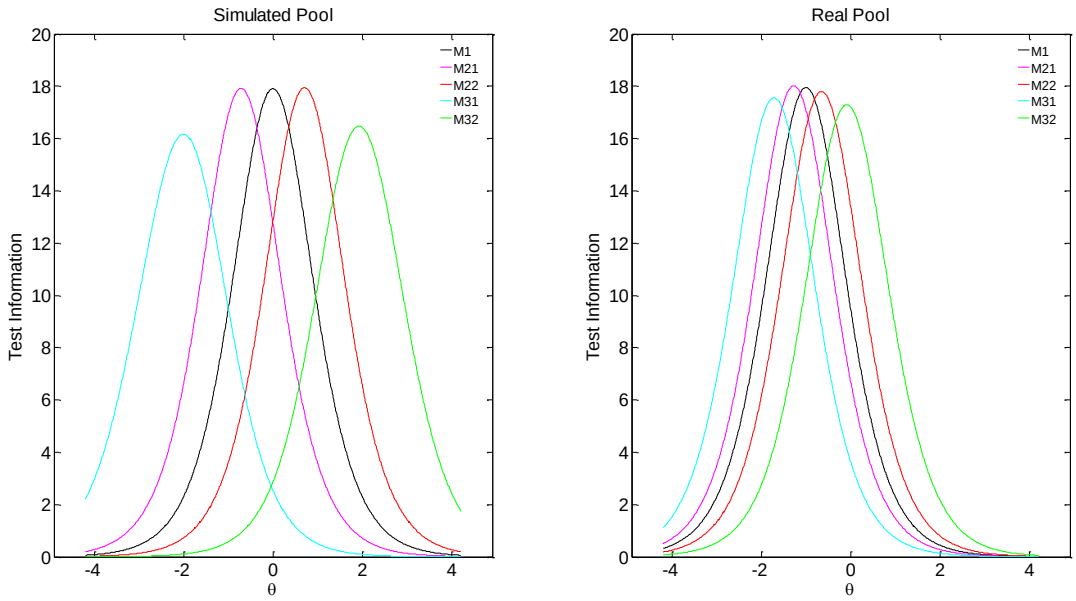

Figure 4.49 Module information functions of the simulated pool and real pool

Table 4.19 to Table 4.20 display the descriptive statistics for both the *S*-Pool and *R*-Pool for the MST 1-2-2 designs under no exposure control and exposure control conditions.

Table 4.19 Item pool descriptive statsitics of the MST 1-2-2 design without exposure control

Item pool type	Pool Size	Mean	SD	Minimum b	Maximum b
R-Pool	125	-0.95	0.60	-2.82	0.56
S-Pool	125	-0.01	1.37	-2.91	2.91

Note: *R*-Pool=real pool; *S*-Pool=Simulated pool

Table 4.20 Item pool descriptive statsitics of the MST 1-2-2 design with exposure control

Item pool type	Pool Size	Mean	SD	Minimum b	Maximum b
R-Pool	1000	-0.96	0.59	-2.82	1.03
S-Pool	1000	-0.02	1.36	-2.97	2.96

Note: *R*-Pool=real pool; *S*-Pool=Simulated pool

The results indicated that the item pool size under exposure control conditions was 8 times larger than those with no exposure control. The mean and standard deviations of the item

pool parameters were quite similar. Since the calibrated item parameters from the real pool were transformed from the logit model scale (e.g., using WINSTEPS) to the probit model scale (e.g., the one used in the present study), the variance of the item parameters was smaller than 1. Table 4.21 to Table 4.22 show the evaluation results by the overall sample for both the *R*-Pool and *S*-Pool under no exposure control and with exposure control conditions. The results indicated that the correlations between true latent ability and estimated ability were both very high for the two types of item pools. The overall bias, RMSE and SE were both quite small and the test reliability and classification accuracy were both high. The marginal reliability for both types of items pools exceeded .95. These results applied to both conditions for without exposure control and with exposure control. When no exposure control was implemented, the item overlap rates for all the conditions across the two item pools were quite large. Under exposure control condition, the item overlap rates all became quite small and within an ideal range of .02 to .20. While comparing the two item pools, it was discovered that the correlations between true ability and estimated ability for the *S*-Pool were slightly higher than those for the *R*-Pool. Under both conditions, the overall bias, RMSE, SE, item overlap rate, and exposure rate for the *S*-Pool were slightly lower than the *R*-Pool, and the overall correlation between true ability and estimated ability, overall reliability, test information, and classification accuracy were slightly higher than the *R*-Pool. The item exposure rate (e.g., conditional on bin) for the simulated item pool was slightly lower than that for the real pool under both exposure control and non-exposure control conditions. Generally speaking, both item pools performed slightly better under exposure control conditions than no exposure control conditions. Under no exposure control condition, no item had exposure rate larger than .20 within each bin. Under exposure control condition, no item had exposure rate larger than .20 and smaller than .02.

Table 4.21 The performance of the MST 1-2-2 optimal item pool without exposure control

Proportion	Corr	Bias	RMSE	Infor	SE	M-Reliability	Classi-fication	Overlap rate	Exposure rate
R-Pool	0.99	0.05	0.27	29.05	0.19	0.96	0.93	0.67	0.55
S-Pool	1.00	0.00	0.18	33.02	0.17	0.97	0.94	0.67	0.54

Note.*R*-Pool=real pool; *S*-Pool=Simulated pool; Corr=Correlation; Infor=Information; SE=standard error of measurement; M-Reliability=marginal reliability

Table 4.22 The performance of the MST 1-2-2 optimal item pool with exposure control

Proportion	Corr	Bias	RMSE	Infor	SE	M-Reliability	Classi-fication	Overlap rate	Exposure rate
R-Pool	0.99	0.05	0.28	28.41	0.19	0.96	0.93	0.08	0.07
S-Pool	1.00	0.00	0.18	33.04	0.17	0.97	0.95	0.08	0.07

Note.*R*-Pool=real pool; *S*-Pool=Simulated pool; Corr=Correlation; Infor=Information; SE=standard error of measurement; M-Reliability=marginal reliability

Figure 4.50 and Figure 4.51 display the results for classification accuracy for the two types of item pools. The cutoff score is determined by the real passing rate of the operational licensure exam used in the study. It is observed from the figure that the *S*-Pool and *R*-Pool performed similarly well in the cutoff score under both exposure control and no exposure control conditions. The *S*-Pool performed slightly better than the *R*-Pool under these two conditions.

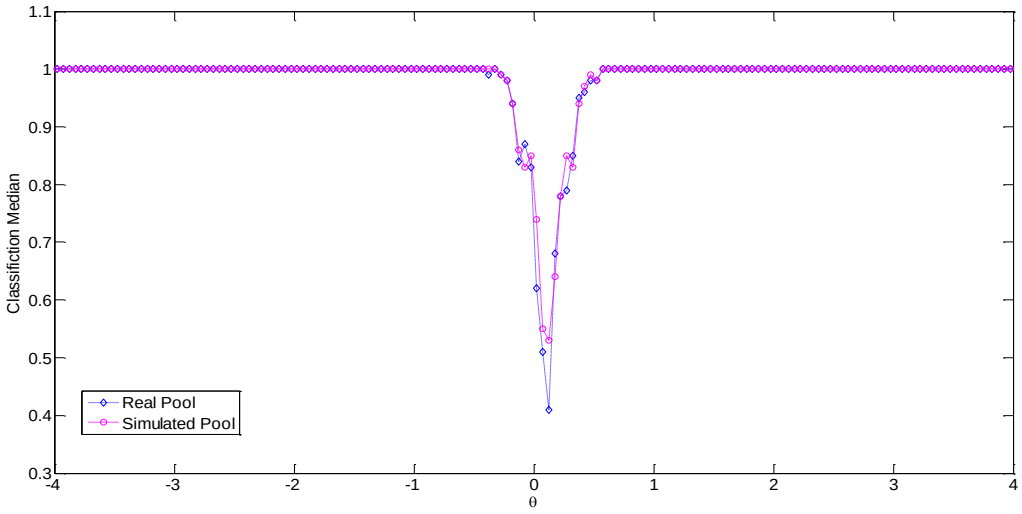

Figure 4.50 Classification accuracy of the real pool and simulated pool without exposure control

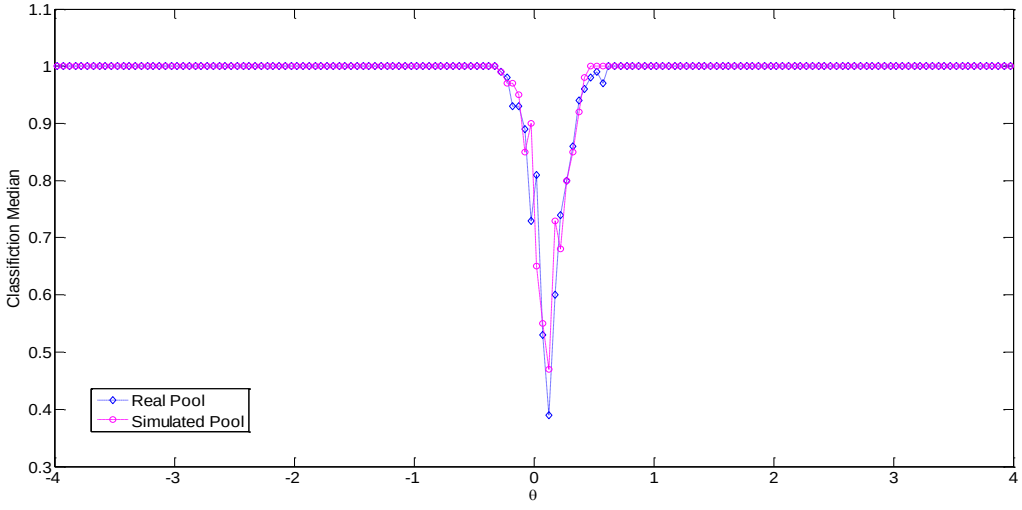

Figure 4.51 Classification accuracy of the real pool and simulated pool with exposure control

Figure 4.52 and Figure 4.53 show the results of the conditional bias displayed by the two item pools. The conditional bias for the *S*-Pool was quite low along the whole range of the θ scale, and the conditional bias for the *R*-Pool was low between -4 to 1 on the θ scale. The same results were found for conditional RMSE (See Figure 4.54 and Figure 4.55) and conditional SE for the two item pools (See Figure 4.56 and Figure 4.57). The *S*-Pool was slightly lower in terms of the conditional SE at the lower end of the ability scale than the *R*-Pool. It should be noted that the decreased conditional bias and RMSE at the higher end of the ability scale for the *R*-Pool is artifact because when a maximum ability is assigned on the ability scale, the conditional bias and RMSE tend to be zero. Thus the end part of the conditional bias and RMSE for the *R*-Pool may be ignored. These results applied to both no exposure control and with exposure control conditions.

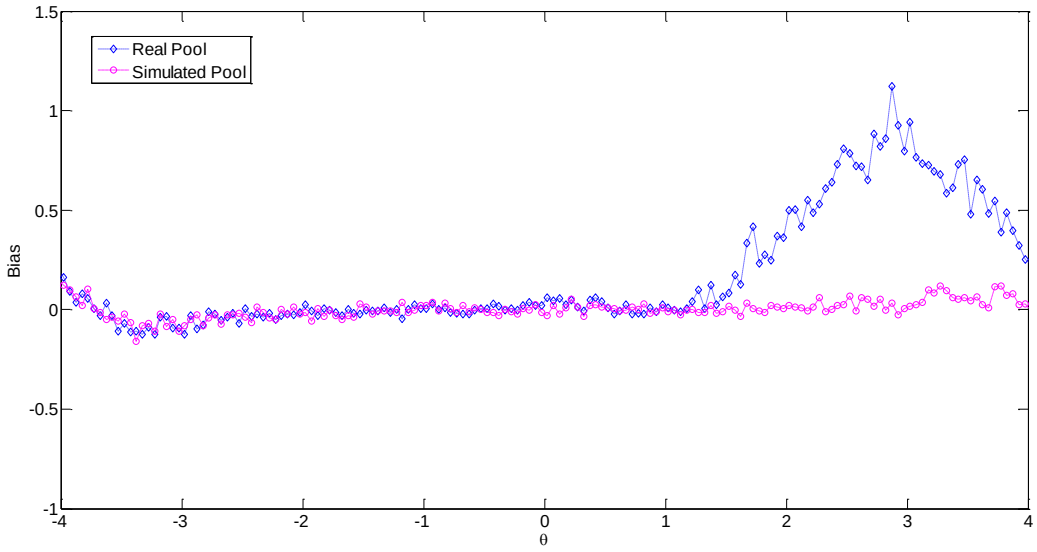

Figure 4.52 Conditional bias of the real pool and simulated pool without exposure control

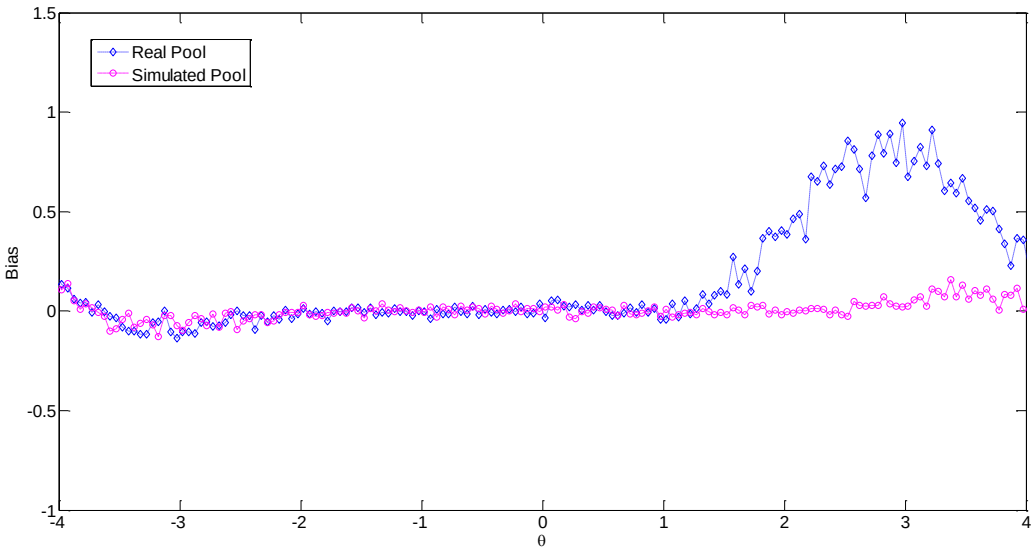

Figure 4.53 Conditional bias of the real pool and simulated pool with exposure control

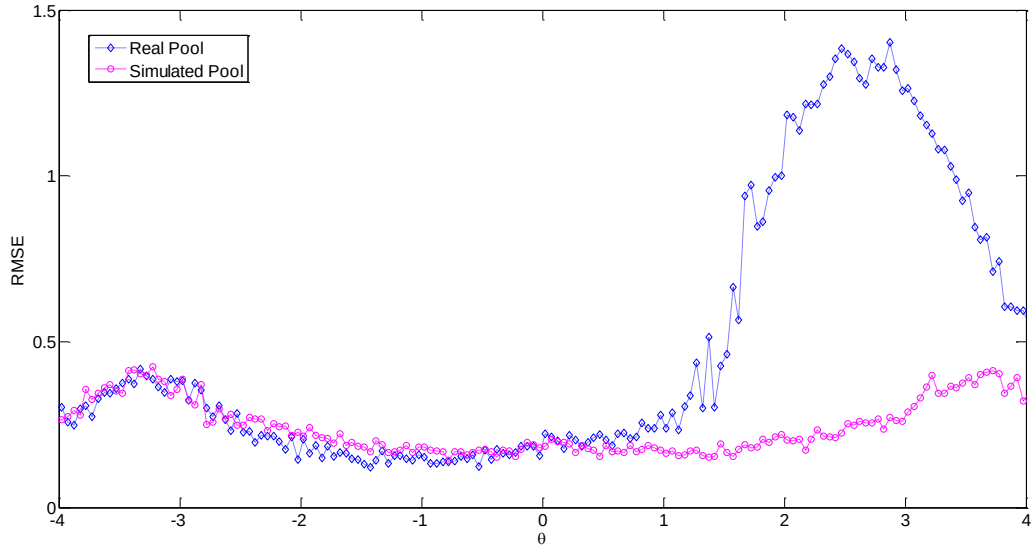

Figure 4.54 Conditional RMSE the real pool and simulated pool without exposure control

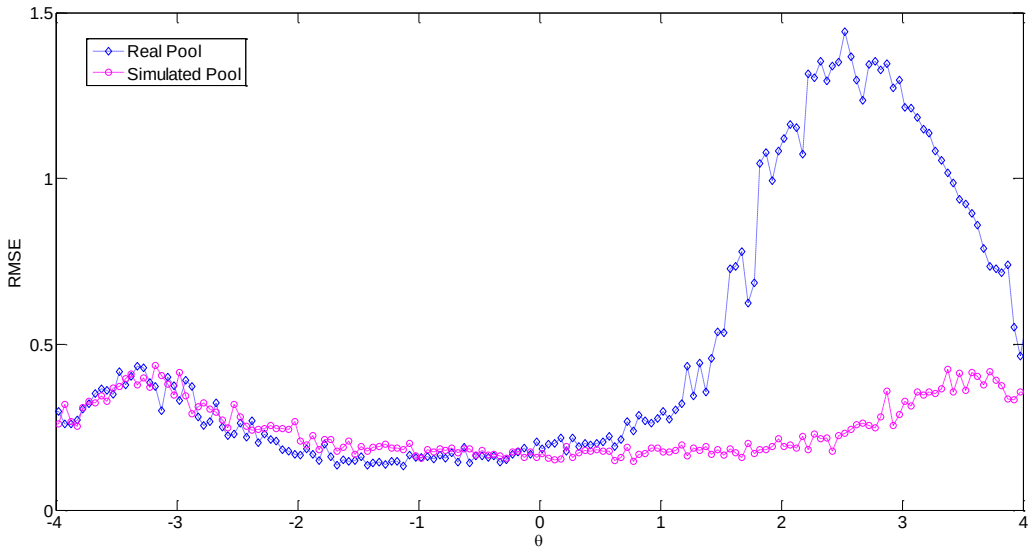

Figure 4.55 Conditional RMSE of the real pool and simulated pool with exposure control

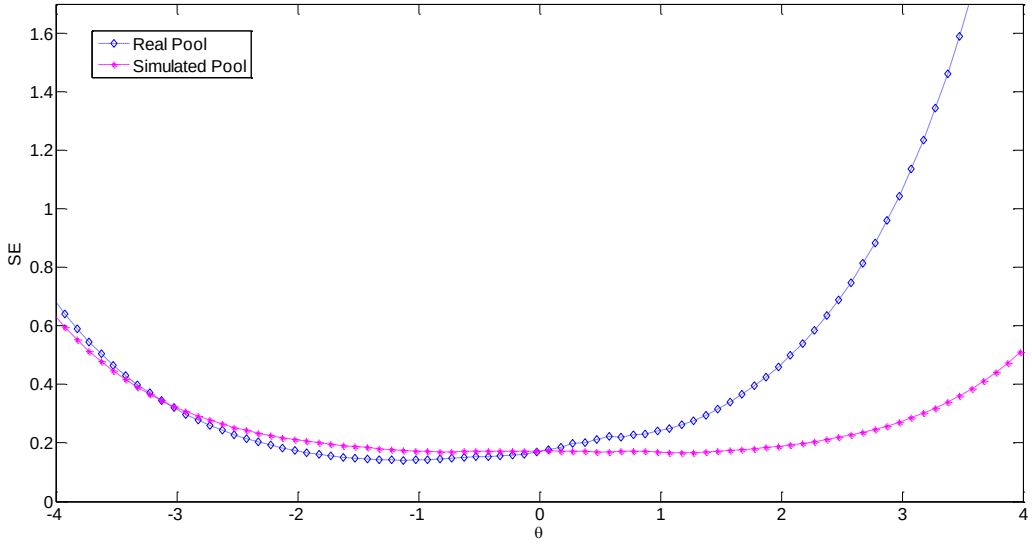

Figure 4.56 Conditional SE of the real pool and simulated pool without exposure control

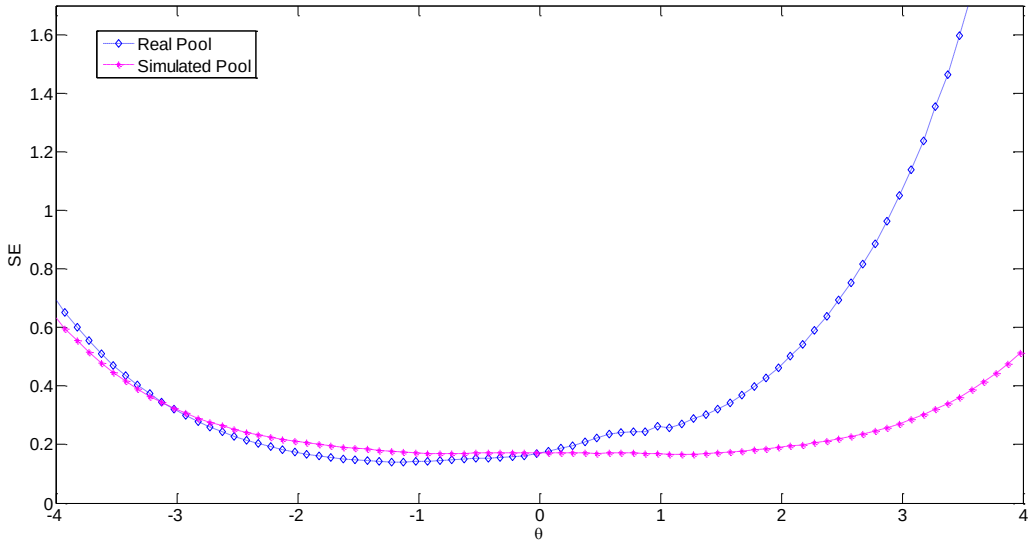

Figure 4.57 Conditional SE of the real pool and simulated pool with exposure control

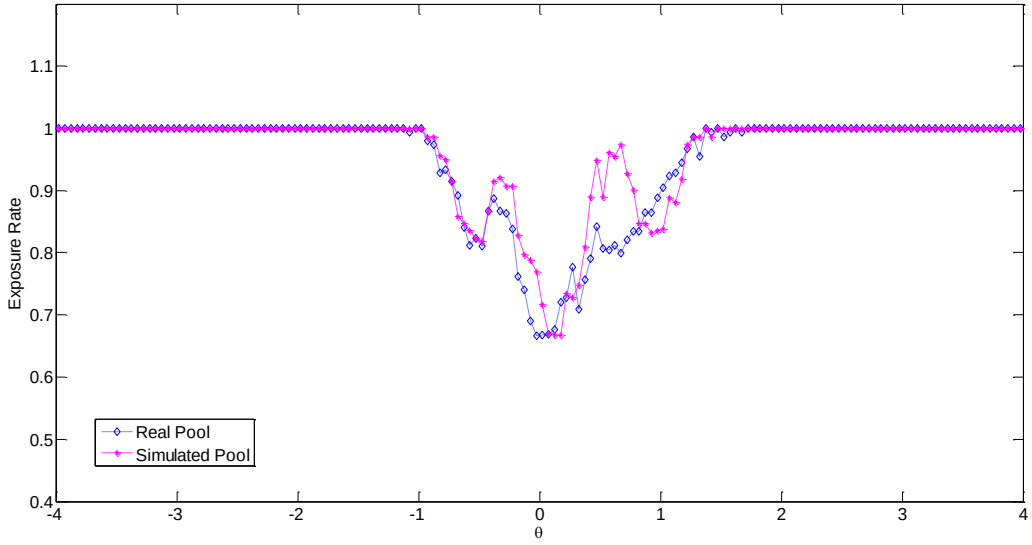

Figure 4.58 Conditional item overlap rate of the real pool and simulated pool without exposure control

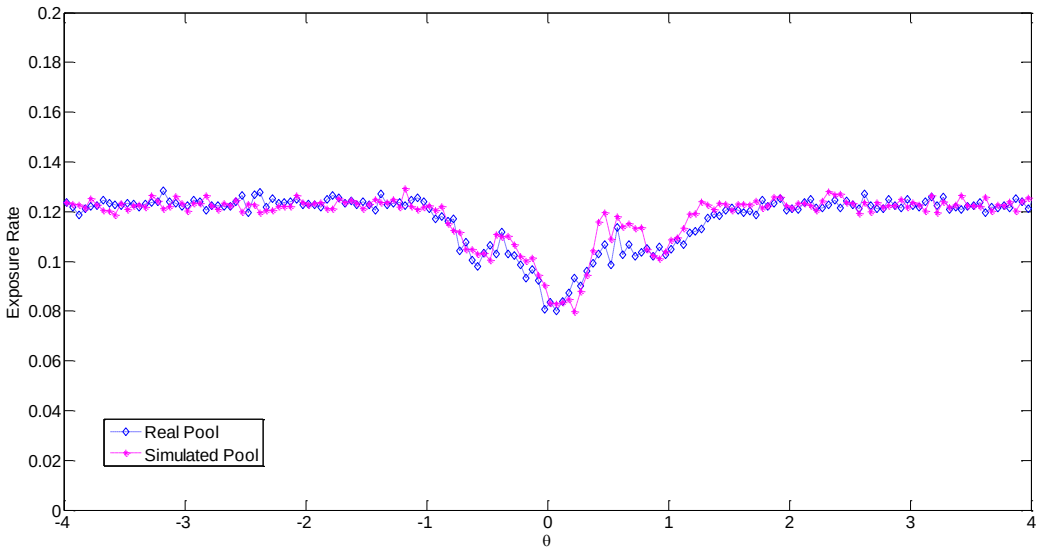

Figure 4.59 Conditional item overlap rate of the real pool and simulated pool with exposure control

Figure 4.58 and Figure 4.59 show the conditional overlap rates for both without and with exposure control conditions. It is concluded from the figures that the items in the middle had lower overlap rate than those at the extremes for both situations. Since there is only one test form being administered, the conditional overlap rate was very high when no exposure control was implemented. After the exposure control was implemented, the conditional overlap rates for both simulated and real item pools fell within the range of .02 to .20.

CHAPTER 5: Discussion

This chapter first summarizes the findings of the study and then discusses the results and implications from the study. Recommendations on how to apply the *p*-optimality method to an operational item pool are discussed. Finally, limitations and future research recommendations are presented.

5.1 Summary of the results

The study was designed to use optimal item pools to support different MST panel designs by extending the *p*-optimality method in Reckase (2003, 2010), and see how the change of test length and routing test length impacted measurement accuracy. Seventy-two optimal item pools were simulated with and without the practical constraint of exposure control, and comparison between simulated and operational item pool performances was also conducted.

In the simulated item pools, the item frequency distributions across different MST designs and test configurations did not have any uniform characteristics. Generally speaking, some item distributions were symmetric and some were skewed. More items were distributed at the value of the cutoff score especially when the routing test proportions were large, such as 40%. The item distributions became more peaked with the increase in routing test proportions. Under the conditions of without and with exposure control, the distributions of item parameters and item pool characteristics for all the MST designs were similar. One exception was the item pool size. Depending on the types of MST designs and routing test proportions, the item pool sizes under the exposure control conditions were about 7 to 11 times larger than those under no exposure control.

The test information functions in the test development stage for all the MST designs revealed that most of the information distributions were symmetric and the information curves

exceeded the required test information. These results applied to the situations for both with and without exposure control. The module information curves in Figure 4.26 for all MST designs displayed that modules built using the *p*-optimality method were parallel with each other. These results all support the fact that the item pool parameters selected using the *p*-optimality method were appropriate to support the different MST panel designs.

The item pool performance by the overall sample indicated that the examinees' true latent ability and estimated ability were all highly correlated for all test configurations across all MST designs. The overall bias and RMSE were all quite small. With increase of test length, the RMSE and standard error of measurement decreased, and the overall test information, reliability and classification accuracy increased. In terms of item pool usage, under the exposure control conditions, the item overlap rates for all the items were between .02 to .20, and the same was true with item exposure rate conditional on bin. Under the no exposure control condition, no item had an exposure rate larger than .20 within each bin. The test length also impacted the classification accuracy. With increase of test length, there was a tendency for more accurate classifications regarding the median, minimum competence and scholarship cutoff scores. However, no obvious difference in classification accuracy was observed when the routing test length varied within a given overall test length. These results applied to those under both exposure and no exposure conditions.

Conditional evaluation results also showed that test length impacted measurement accuracy for all MST designs. With increase of test length, the conditional bias, RMSE and SE all decreased. The bias was the lowest when the test length is 60 and the highest when the test length is 20, especially at the two extremes where negative bias was found at the lower end and positive bias was found at the higher end. The conditional bias, RMSE and SE were the highest

when the routing test proportion is 20%, however, not much difference was found between 30% and 40% routing test proportions. These results applied to all the exposure control and no exposure control conditions.

From the results, it is concluded that all the MST designs worked well in achieving high measurement accuracy regarding correct classification accuracy. Some variation arose with the change of test length and routing test length. Generally speaking, among all the designs, based on the 95% correct classification rates, the MST 1-2-3 design worked the best in classification accuracy for a median cutoff score for a test with 20 and 40 items for all routing test proportions except for the 30% proportion; for the minimum competence cutoff score for a test of 60 items when the routing test proportion is 20% and for a test of 40 items when the routing proportions are 30% and 40%; and for the scholarship cutoff score for a test of 60 items when the routing test proportion is 20% and 30%. The MST 1-3 design worked the best for the minimum competence and scholarship cutoff scores for a test of 20 items with all routing test proportions; for a test of 40 items for the minimum competence and scholarship cutoff scores when the routing test proportion is 20%; for a test length of 60 items for the median cutoff score where the routing test proportion is 40%; and for the scholarship cutoff score when the routing test proportion is 40%. The MST 1-2-2 design worked the best for a test length of 40 for the minimum competence and scholarship cutoff scores when the routing proportions are 20% and 30%, and for a test length of 60 for the minimum competence cutoff score when the routing test proportion is 30%. The MST 1-2 design had the highest classification accuracy for a test length of 40 for the scholarship cutoff score when the routing proportion is 40%; and for a test length of 60 for the minimum competence cutoff score when the routing proportion is 40%.

Regarding the conditional evaluation results as shown by the conditional bias, RMSE, SE and item overlap rate, not much difference was discovered across the different MST designs. When the test length is 20, the MST 1-2-2 design had slightly higher bias, RMSE and SE than the rest of the designs, especially at the two extremes on the θ scale. The MST 1-2-3 design had much lower conditional item overlap rate than the rest of the three designs across the whole θ scale, but not much difference was discovered across the rest of the three MST designs. The values of the conditional statistics, such as the conditional bias, RMSE and SE were found to decrease with the increase of test length. From the comparisons between the results under exposure control and no exposure control conditions, it is concluded that the item pool sizes under exposure control procedures were about 7 to 11 times larger than those without implementing the exposure control. The means and standard deviations of the item pool parameters were quite similar under the two conditions.

The application of the *p*-optimality method in an operational item pool in a multistage licensure adaptive test indicated that this method was feasible in item pool design under real testing situations. The simulated optimal item pool had almost as good classification accuracy as the real pool. Compared with the real pool, the simulated pool covered the whole range of the ability scale in terms of good measurement accuracy. The measurement accuracy for the real pool was insufficient in the range above 1 on the ability scale. For conditional exposure rates and item overlap rates under exposure control conditions, both the real pool and simulated pool fell between the ideal range of .02 to .20.

5.2 Discussion of the results

Van der Linden et al. (2006) defined the optimal item pool as "consisting of a maximal number of combinations of items that (a) meet all content specifications for the test and (b) are

most informative at a series of ability levels reflecting the shape of the distribution of the ability estimates for a population of examinees" (p. 82).

As indicated in the results in both the simulation study and the empirical study, the simulated item pools built by the *p*-optimality method achieved good measurement accuracy along the whole range of the θ scale for all MST designs. This result is consistent with the optimal item pool design study in CAT by Reckase (2003, 2010), Gu (2007), He and Reckase (2014), Zhou (2012) and Mao (2014). This method was also proved to be effective in designing the optimal item pool for an operational licensure exam in multistage adaptive test.

Although all the MST designs achieved similarly good measurement accuracy along the whole range of the θ scale, slight differences were discovered for different types of cutoff scores with different test lengths and routing test lengths. For example, the MST 1-2-2 design worked the best for the minimum competence and scholarship cutoff scores when the routing proportions are 20% and 30% for a 40-item test. However, the MST 1-2-3 design achieved better measurement accuracy than the rest of the designs for the minimum competence cutoff score when the test is composed of 60 items when the routing test proportion is 20% and of 40 items when the routing proportions are 30% and 40%; and for the scholarship cutoff score when the test is composed of 60 items when the routing test proportion is 20% and 30%. Therefore, in operational testing, the choice of the best MST design is dependent on multiple factors, such as test length, test configuration and test purpose.

Not much difference was discovered regarding accurate ability estimation across the four MST designs in the present study given a certain test length and test configuration. This result is not consistent with those discussed in Patsula (1999), which concluded that three-stage MSTs produced smaller measurement error than two-stage MSTs by comparing four MST test

structures 1-3, 1-5, 1-3-3, and 1-5-5. This also confirms previous arguments about the choice of the number of stages for MST design, which stated that increasing the number of stages did not necessarily bring more measurement precision. (Jodoin et al., 2006). Similarly, the variation of the routing test proportions also did not show much difference in the results across all the MST designs under all test specifications. This result is consistent with the previous research about the importance of the routing test length in MST design studies (Patsula, 1999; Zheng et al., 2012). However, the test length did impact the performances of all the MST designs. The overall and conditional bias, RMSE and SE all decreased with the increase of test length.

Comparing the no exposure and with exposure control conditions, depending on the type of MST design and test length as well as routing test length, the latter had an item pool size that was about 7 to 11 times larger than the former. The previous study found a situation where the implementation of exposure control enlarged the item pool size by threefold (He & Reckase, 2014). The size of the item pool after the implementation of exposure control is the result of considering multiple factors, such as the routing test length, number in the examinee population, test mode and test designs. As revealed in the study results, the measurement accuracy was almost equally good for before and after implementing the exposure control. For the exposure control conditions, however, the measurement accuracy was higher and the conditional overlap rates all fell between an ideal range of .02 to .20. This implies that no items were overexposed or underexposed and the item pool usage was appropriate. This result applied to the two types of exposure control conditions implemented in the simulation study and empirical study.

The inverse proportion exposure control method is advised for future use in operational testing since it achieved good measurement accuracy for all MST designs. Comparatively speaking, for the exposure control method where an equal number of test forms were assembled

for modules at each stage, some items were actually not used sufficiently, especially for the second and third stage modules. Since item creation is very costly, all items are expected to be sufficiently used in operational testing administrations. The simulated item pool by the *p*-optimality method covered the whole range of the ability scale in terms of good measurement accuracy, the conditional item overlap rates (e.g., under exposure control conditions), and the item exposure rates conditional on bins all fell within the ideal range of .02 to .20, it is applicable in designing optimal item pools for both achievement test and licensure test in operational MST context.

In the MST 1-2-2 and MST 1-2-3 designs, the items in the second stage were all piled up in the bins where the cutoff scores were located. This type of optimal design could reduce the complexity in the item pool design process and is desirable in reality. As done in the first stage, the items could be created around the cutoff scores in the second stage where maximum information is obtained for the provisional ability estimates of the examinees. There are several advantages for this type of design. First, given the target population distribution, the cutoff scores at the easy module and hard module in the second stage could be easily located before test administration and be implemented in the testing procedures. Second, it is more likely that misrouting is reduced to the minimum when the items at the second stage are all centered around the cutoff scores. Third, with this design, an equal proportion of the examinees could be routed to the third stage based on the two cutoff scores at the second stage. Fourth, a large proportion of the items, as much as 40% of the total test, are intentionally designed for the third stage in order to ensure that sufficient measurement accuracy are achieved. Instead, suppose the items at the second stage were scattered across bins and the two cutoff scores were calculated in the testing process based on the estimated ability characteristics, part of the measurement accuracy would

be lost in order to compensate the decreased classification accuracy after the second stage test. Future studies might be conducted to see how much measurement accuracy could be lost if the items selected were scattered across the whole range of the ability scale at the second stage for MST 1-2-2 and MST 1-2-3 designs.

5.3 Implications from the study

The output of the *p*-optimality method is a blueprint for designing the optimal item pool of the multistage computerized adaptive test. The number of items within each bin and the bin locations along the θ-scale provide the guidance for the statistical attributes that the items are supposed to possess in the process of item writing and test assembly. The non-statistical attributes, such as content balancing, can also be achieved through a built-in design. For example, the bins are designed for each of the anticipated content areas and then are combined together. Due to unidimensionality assumption and the implementation of a hypothetical test in the study, content balancing was not addressed. Future studies might consider choosing a specific operational test and implement the content balancing in the process of applying the *p*-optimality method in the optimal item pool design.

As revealed in the different information functions in the item pool design stage, the shapes of the information functions across the different MST designs under different test configurations were slightly different. This implies that items selected for different test configurations and test purposes should be different. If the test is designed to select examinees for scholarships, the examinee distribution for simulation procedures could follow a negatively skewed distribution and more items are selected to get higher measurement precision at the higher end. Essentially, the optimal item pool designed by the *p*-optimality method is to achieve a high measurement precision for a wide range of abilities along the θ-scale. Therefore, with a

normally distributed examinee population, the most popular distribution in reality, it could achieve an ideal measurement precision at all kinds of cutoff scores for different test purposes. The comparison results from the study about different MST designs and different test configurations could provide some general references for the optimal item pool design for future operational MSTs.

Considering which MST design under which test configuration is the best in achieving the best measurement precision, there is not a uniform answer. Different MST designs perform differently given different conditions and cutoff scores, but all achieve desirable measurement accuracy along the whole θ-scale. The same is true with the operational item pool study. Therefore, from the cost-effective perspective, one could choose the simplest design, such as MST 1-2 design, to keep the cost of item creation and maintenance down. The MST 1-3 design is also good for consideration due to its simplicity in the whole testing algorithm and accurate classification accuracy as displayed in the results under several conditions. If for any reason the test practitioners hope to choose one of the three-stage designs, both the MST 1-2-2 and MST 1-2-3 designs are good choices. The former is advantageous in that it requires fewer items than the latter. The latter has slightly higher measurement accuracy than the former and can be considered for various tests designed for various measurement purposes, such as for pass/fail decision, selection of minimum competent or scholarship recipients.

Test length is an important factor to consider while designing a test for MST. As is seen from the present study, a test length of 20 is not long enough to accomplish all the MST designs with a good measurement precision, especially for MST 1-2-2 and MST 1-2-3 design. Comparatively speaking, the test lengths of 40 and 60 are more appropriate for future MST designs. Regarding the routing test proportions, although the variation among them did not bring

much difference in the results, 20% is not recommended since it deviated from the 30% and 40% conditions in terms of all evaluation statistics. The 30% and 40% conditions did not vary a lot from each other. Given this situation, with the purpose of accumulating more items at the final stage for a better measurement precision, a routing test proportion of 30% is recommended for future test use.

The comparison results between the real pool and simulated optimal item pool using the *p*-optimality method indicates that the method is applicable in real testing situations. However, if the item parameters in the real pool do not align well with the characteristics of the target population, direct application of the method may cause higher measurement error along the θ-scale where the non-alignment occurs. Therefore, in real operational testing situations, the target examinee population characteristics need to be followed while designing the optimal item pool to support the various MST panel designs.

Item pools are not a static entity. Some items become obsolete after they are released to the public or overexposed due to repeated test administrations. New items are frequently added to the item pools on a continuous basis and new item pools are established. van der Linden et al. (2000) discussed the difficulties of item pool management using the integer programming approach. By this approach, a previous item pool was needed to define a cost function and design the new item pool with newly added items. It was anticipated that the newly added items should address the same attributes of the items in the previous item pool so that the cost function could perform appropriately. However, as stated in van der Linden et al. (2000), it is possible that the attributes of the old items in the previous pool do not constitute a practical solution because they do not address the same attributes in the new items. Therefore, if new attributes of the items are introduced, this approach "might face an unsolvable missing data problem" (van der Linden et

al., 2000, p. 148). Comparatively speaking, the optimal item pool designed by the *p*-optimality method can be easily adapted to fit for the new blueprint and maintain the appropriate item pool development and management. More specifically, the retired items can be replaced by newly created items in the same bins since they share the same psychometric attributes.

5.4 Limitations and future recommendations

The results of the study demonstrated the advantages of the optimal item pools designed by the *p*-optimality method to support the different MST designs in achieving good measurement precision and item pool usage. However, this conclusion is restricted by several factors. For example, the items are assumed to be fit by the Rasch model under the unidimensionality assumption. As we know, Rasch model requires the data to fit the model. The values of misfit need to be considered before evaluating the final calibration results. Future research might consider using other IRT models, such as 2PL and 3PL models and compare the results.

Based on the unidimensionality assumption and the nature of the hypothetical test in the simulation study, content balancing was not considered and items from different content areas were assumed to have the same distributions. Although previous research concluded that content balancing had little impact on the measurement precision under different conditions (He & Reckase, 2014; Zhou, 2012), it could be implemented in future studies with operational tests and see how the results actually work in practice. Multidimensional IRT models are also suggested in future studies and see how optimal item pools are designed for operational MSTs where multiple latent traits or subscores are considered and reported.

In addition, the exploration of the test lengths and routing test proportions in MSTs were not exhaustive and they only represent short, medium and long test length and routing test length. Future studies might consider using other test lengths and routing test proportions as

needed and extend the results from the study. Furthermore, instead of using the fixed-length test in MST design, future studies might consider applying the variable-length MST design. For example, suppose examinees achieve very high scores or very low scores after the last stage of the test, and the measurement precision near the cutoff score is still not very high, an additional module might be administered so that they have some chance to recover from the possible accidental misrouting.

The item types in the study are restricted to dichotomous items. As we know, it is possible that the MSTs are composed of different item types, such as polytomous items, testlet-based items, performance-based items, and mixed-format items. It is of interest if future studies consider incorporating a different item type and examine how the *p*-optimality method applies in that context to support the operational MST assembly.

APPENDIX

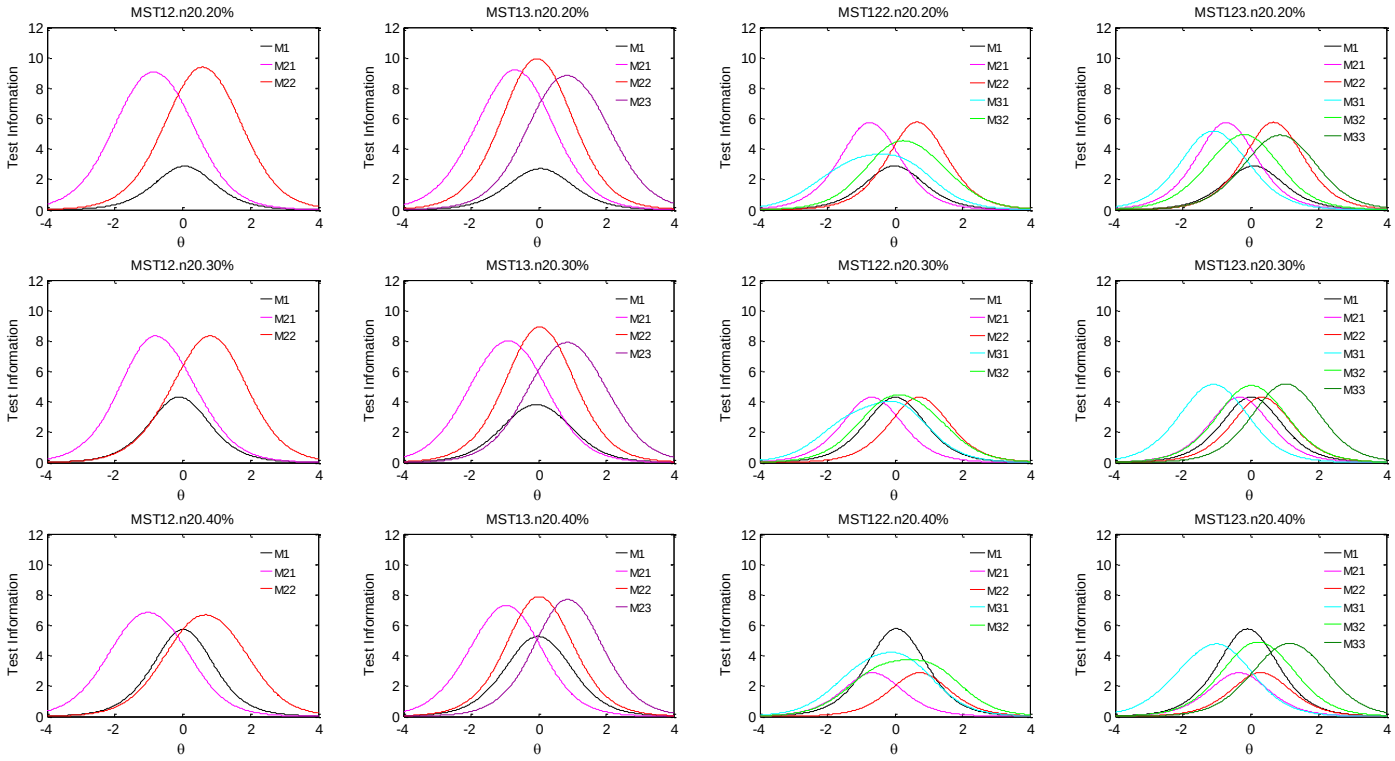

Figure A.1 Module information curves for all test configurations in all MST designs for the test length of 20

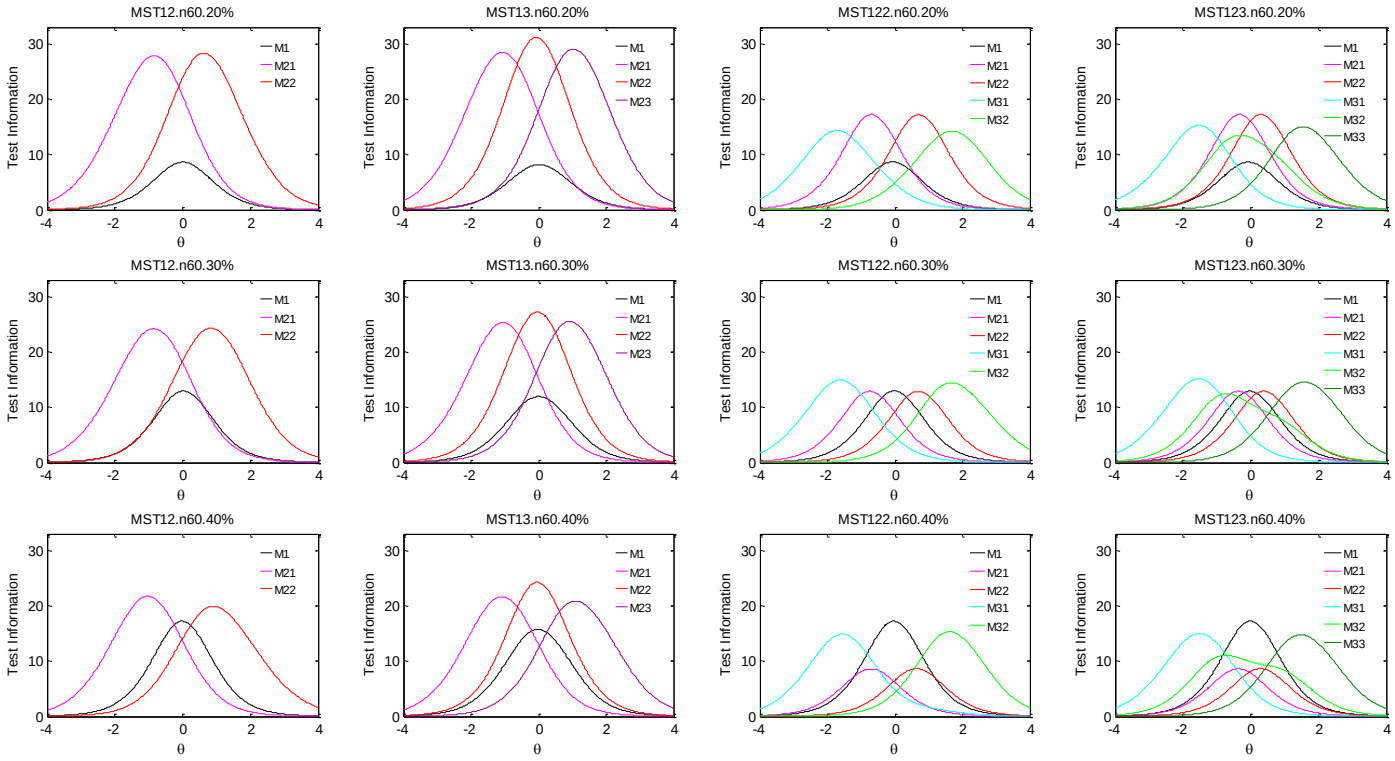

Figure A.2 Module information curves for all test configurations in all MST designs for the test length of 60

REFERENCES

REFERENCES

Adema, J. J. (1990). The construction of customized two-stage tests. *Journal of Educational Measurement, 27*(3), 241-253.

Ariel, A., Veldkamp, B. P., & Breithaupt, K. (2006). Optimal Testlet Pool Assembly for Multistage Testing Design. *Applied Psychological Measurement*, 30(3), 204-215.

Armstrong, R., & Edmonds, J. (2004). *A study of multiple stage adaptive test designs*. Paper presented at the Annual Meeting the National Council on Measurement in Education (NCME), San Diego: CA.

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.

Berger, M. P. F. (1994). A general approach to algorithmic design of fixed-form tests, adaptive tests, and testlets. *Applied Psychological Measurement, 18,* 141–153.

Boekkooi-Timminga, E. (1991). *A method for designing Rasch model based item banks.* Paper presented at the annual meeting of the Psychometric Society, Princeton NJ.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental scores* (pp. 397-479). Reading MA: Addison-Wesley.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of the EM algorithm. *Psychometrika, 46*, 443-459.

Bock, R. D., & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika, 35,* 179-198.

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

Chang, H. H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement, 20*(3), 213-229.

Chen, H., Yamamoto, K., & von Davier, M. (2014). Controlling multistage testing exposure rates in international large-scale assessments. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized Multistage Testing: Theory and Applications* (pp. 391-408). New York: CRC Press, Taylor & Francis Group.

Davis, L. L., & Dodd, B. G. (2003). Item exposure constraints for testlets in the verbal reasoning section of the MCAT. *Applied Psychological Measurement, 27*(5), 335-356.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York: The Guilford Press.

Embretson, S. E. & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Gu, L. (2007). *Designing optimal item pools for computerized adaptive tests with exposure controls*. Unpublished doctoral dissertation. Michigan State University.

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, W., & Diao, Q. (2014). *Item Pool Design for CAT-Review, Demonstration, and Future Prospects*. Paper presented at the Annual Meeting of the National Council on Measurement in Education (NCME), Philadelphia, PA.

He, W., & Reckase, M. (2014). Item pool design for an operational variable-length computerized adaptive test. *Educational and Psychological Measurement, 74*(3), 473-494.

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

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.

Keng, L. (2008). *A comparison of the performance of testlet-based computer adaptive test and multistage tests.* Unpublished doctoral dissertation, University of Texas,

Kim, J., Tseng, C. H., Chung, H., & Dodd, B. G. (2008). *A comparison of the test design variations in panel structures of the computerized adaptive sequential testing system under the partial credit model.* Paper presented at the annual meeting of the American Educational Research Association, New York City, NY.

Kim1, J., Chung, H., Dodd, B. G., & Park, R. (2012). Panel Design Variations in the Multistage Test Using the Mixed-Format Tests, *Educational and Psychological Measurement, 72*(4), 574–588.

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

Luecht, R. M. (1998). Computer-assisted test assembly using optimization heuristics. *Applied Psychological Measurement, 22*(3), 224-236.

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

Luecht, R. & Nungester, R. (1998). Some Practical Examples of Computer-Adaptive Sequential Testing. *Journal of Educational Measurement, 35*(3), 229–249.

Luecht, R. M., & Nungester, R. J. (2000). Computer-adaptive sequential testing. In C. Glas & W. J. van der Linden (Eds.), *Computer-Adaptive Testing* (pp. 117-128). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Luecht, R. M., Brumfield, T., & Breithaupt, K. (2002). *A testlet assembly design for the uniform CPA examination.* Technical Report: Series Two.

Luecht, R. (2003). *Exposure control using adaptive multi-stage item bundles*. Paper presented at the Annual Conference of Meeting of the National Council on Measurement in Education (NCME), Chicago, IL.

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., & Hadidi, A. (1996). *Heuristic-based CAT: Balancing item information, content and exposure*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New York.

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

Melican, G. J., Breithaupt, K. & Zhang, Y. (2010). Designing and implementing an multistage adaptive test: The Uniform CPA Exam. In van der Linden et al. (Eds.) *Elements of Adaptive Testing*, (pp. 167-189). New York: 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.

Park, R. (2013). *The impact of Statistical constraints on classification accuracy for Multistage tests*. The Uniform CPA Exam technical report.

Patsula, L. N. (1999). *A comparison of computerized adaptive testing and multistage testing*. Unpublished Doctoral dissertation. University of Massachusetts Amherst, MA.

Patsula, L. N., & Hambleton, R. K. (1999). *A comparative study of ability estimates obtained from computer-adaptive and multi-stage testing*. Paper presented at the Annual Conference of the National Council on Measurement in Education (NCME), Montreal, Quebec, Canada.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment test*. Copenhagen: The Danish Institute for Educational Research.

Reckase, M. D. (2003). *Item pool design for computerized adaptive tests*. Paper presented at the Annual Meeting of the National Council on Measurement in Education (NCME), Chicago, IL.

Reckase, M. D. (2006). *Design of an ideal two-stage test*. Paper presented at the Annual Meeting of the National Council on Measurement in Education (NCME), San Francisco, CA.

Reckase, M. D. (2010). Designing item pools to optimize the functioning of a computerized adaptive test. *Psychological Test and Assessment Modeling*, *52*(2), 127-141.

Rotou, O., Patsula, L., Manfred, S., & Rizavi, S. (2003). *Comparison of Multi-stage Tests with Computerized Adaptive and Paper and Pencil Tests.* Paper presented at the annual meeting of the American Educational Research Association (AERA) and the National Council on Measurement in Education (NCME), Chicago, IL.

Segall, D. O., Moreno, K. E., & Hetter, D. H. (1997). *Item pool development and evaluation*. In W. A. Sands, B. K. Waters, & J.R. McBride (Eds.), Computerized adaptive testing: From inquiry to operation (pp. 117–130). Washington DC: American Psychological Association.

Stocking, M.L, & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement, 17,* 151-166.

Thissen, D. (2000). Reliability and measurement precision. In H. Wainer (Ed), *Computerized adaptive testing: A primer* (2nd ed., pp. 159-183). Mahwah, NJ: Lawrence Erlbaum Associates.

Timminga, E., van der Linden, W. J., & Schweizer, D. A. (1996). ConTEST [Computer program and manual]. Groningen, The Netherlands: iecProGAMMA.

van der Linden, W. J., & Boekkooi-Timminga, E. (1989). A maximin model for test design with practical constraints. *Psychometrika, 54*(2), 237-247.

van der Linden, W. J., Veldkamp, B. P., and Reese, L. M. (2000). An integer programming approach to item pool design. *Applied Psychological Measurement, 24(2),* 139-150.

van der Linden, W. J. (2005). *Linear Models of Optimal Test Design*. New York: Springer.

van der Linden, W. J., Ariel, A., & Veldkamp, B. P. (2006). Assembling a CAT item pool as a set of linear tests. *Journal of Educational and Behavioral Statistics, 31,* 81-99.

Veldkamp, B. P. (2014). Item pool design and maintaince for multistage testing. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized Multistage Testing: Theory and Applications* (pp. 39-53). New York: CRC Press, Taylor & Francis Group.

Veldkamp, B. P., & van der Linden, W. J. (2000). Designing item pools for computerized adaptive testing. In van der Linden, W.J., & Glas, C.A.W. (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 149-162). The Netherlands: Kluwer Academic Publishers.

Wainer, H., & Kiely, G. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24,* 185-202.

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

Wang, W. C., & Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement, 29,* 126-149.

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

Weissman, A., Belov, D. I., & Armstrong, R. D. (2007). *Information-based versus number-correct routing in multistage classification tests.* LSAC Research Report Series. A publication of the Law School Admission Council.

Weissman, A. (2014). IRT-Based Multistage Testing. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized Multistage Testing: Theory and Applications* (pp. 153-168). New York: CRC Press, Taylor & Francis Group.

Wentzel, C., Mills, C. M., & Meara, K. C. (2014). Transitioning a K-12 Assessment from Linear to Multistage Tests. In Yan, D., von Davier, A. A., & Lewis, C. (Eds.), *Computerized Multistage Testing: Theory and Applications* (pp. 355-369). New York: CRC Press, Taylor & Francis Group.

Winer, B. J., Brown, D. R., & Michels, K. M. (1991). *Statistical Principles In Experimental Design.* New York, NY: McGraw-Hill.

Xing, D., & Hambleton, R. K. (2004). Impact of test design, item quality, and item bank size on the psychometric properties of computer-based credentialing examinations. *Educational and Psychological Measurement, 64*(1), 5-21.

Yan, D., Lewis, C., & von Davier, A. (2014). Overview of computerized multistage tests. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized Multistage Testing: Theory and Applications* (pp. 3-20). New York: CRC Press, Taylor & Francis Group.

Zenisky, A. L. (2004). *Evaluating the effects of several multi-stage testing design variables on selected psychometric outcomes for certification and licensure assessment.* Unpublished doctoral dissertation, University of Massachusetts, Amherst.

Zenisky, A. L., Hambleton, R. K., & Luecht, R. M. (2010). Multistage testing: Issues, designs, and research. In W. J. van der Linden & C.A. W. Glas (Eds). *Elements of adaptive Testing* (pp. 355-372). New York: Springer.

Zhang, Y., Breithaupt, K., Tessema, A., & Chuah, D. (2006). *Empirical vs. Expected IRT-Based Reliability Estimation in Computerized Multistage Testing (MST).* Paper Presented at the Annual Conference of the National Council of Measurement in Education, San Francisco, CA.

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*. Paper presented at the Annual Meeting of the National Council on Measurement in Education (NCME). Vancouver, British Columbia, Canada.

Zhou, X. (2014). *Designing P-optimal item pools in computerized adaptive tests with polytomous items*. Unpublished doctoral dissertation. Michigan State University.