The Dissertation Committee for Ryoungsun Park
certifies that this is the approved version of the following dissertation:

INVESTIGATING THE IMPACT OF A MIXED-FORMAT ITEM POOL ON OPTIMAL TEST DESIGNS FOR MULTISTAGE TESTING

Committee:

Barbara G. Dodd, Supervisor

S. Natasha Beretvas

Tiffany A. Whittaker

Jodi M. Casabianca

Xiaojing Jadie Kong

INVESTIGATING THE IMPACT OF A MIXED-FORMAT ITEM POOL ON OPTIMAL TEST DESIGNS FOR MULTISTAGE TESTING

by

Ryoungsun Park, B.S.E.E.; M.S.E.

DISSERTATION

Presented to the Faculty of the Graduate School of
The University of Texas at Austin
in Partial Fulfillment
of the Requirements
for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN
May 2015

Dedicated to my parents

Acknowledgments

I want to express special appreciation for my advisor Dr. Barbara Dodd, who has supported my journey since the beginning. Without her caring support, I could never have completed this dissertation. She did not merely direct me about what to do during my studies; rather, she also guided and mentored me in many aspects of my program. As a result, I have acquired insights into what it is like to advise a disoriented graduate student like myself. It requires quite amount of patience to listen and offer the best guidance in specific situations. I feel as if I have only just made it to the first milestone in my lifelong voyage, and Dr. Dodd has carried me.

I also want to thank my dissertation committee members, Dr. Natasha Beretvas, Dr. Tiffany Whittaker, Dr. Jodi M. Casabianca, and Dr. Xiaojing Jadie Kong. I appreciate Dr. Beretvas for her positive and constructive feedback. I owe my sincere gratitude to Dr. Whittaker for her gentle guidance when needed. Dr. Casabianca offered wonderful help when I was preparing for my job interview, sitting down with me to go over the details. Last but not least, special thanks go to my friend and mentor Jadie, whose support has gone well beyond this dissertation. Her wisdom and advice from her real-life experiences always served as timely help for me and my family.

I also wish to thank Chen family (i.e., David, Bianca, Jonathan, and

Justin), Bill and Sandra, and Bill and Linda for their prayers over everything concerning me and my family. Their presence made us feel right at home even when we are far away from our families. I also want to thank my mom and dad for cheering me on in my study. Although graduate school has been a long process, they remained patient waiting me to finish the course. I cannot forget to mention my son Andrew! I am so glad you came into our life, creating such a big change for us. Your smile continued to encourage me to go on, and I love you dearly. Finally, my wife Jiseon helped me tremendously during my study program. Her sacrifice, patience, and support were essential for enabling me to finish the work, and I am eternally grateful for everything she has provided for me. Thank you so much for everything. I will never forget the support and help I have received.

INVESTIGATING THE IMPACT OF A MIXED-FORMAT ITEM POOL ON OPTIMAL TEST DESIGNS FOR MULTISTAGE TESTING

Publication No. __________________

Ryoungsun Park, Ph.D.
The University of Texas at Austin, 2015

Supervisor: Barbara G. Dodd

The multistage testing (MST) has drawn increasing attention as a balanced format of adaptive testing that takes advantages of both fully-adaptive computerized adaptive testing (CAT) and paper-and-pencil (P&P) tests. Most previous studies on MST have focused on purely dichotomous or polytomous item formats although the mixture of two item types (i.e., mixed-format) provides desirable psychometric properties by combining the strength of both item types. Given the dearth of studies investigating the characteristics of mixed-format MST, the current study conducted a simulation to identify important design factors impacting the measurement precision of mixed-format MST.

The study considered several factors-namely, total points (40 and 60), MST structures (1-2-2 and 1-3-3), the proportion of polytomous items (10%, 30%, 50% and 70%), and the routing module design (purely dichotomous and a mixture of dichotomous and polytomous items) resulting in 32 total conditions. A total of 100 replications were performed, and 1,000 normally distributed examinees were generated in each replication. The performance of MST was evaluated in terms of the precision of ability estimation across the wide range of the scale.

The study found that the longer test produced greater measurement precision while the 1-3-3 structure performed better than 1-2-2 structure. In addition, a larger proportion of polytomous items resulted in lower measurement precision through the reduced test information during the test construction. The interaction between the large proportion of polytomous items and the purely dichotomous routing module design was identified. Overall, the two factors of test length and the MST structure impacted the ability estimation, whereas the impact of the proportion of polytomous items and routing module design mirrored the item pool characteristic.

Table of Contents

List of Tables

List of Figures

Chapter 1

INTRODUCTION

Item Response Theory (IRT) is the basis for item analysis, test construction, administration, and scoring of many operational tests. With IRT, advancements in computer technology have allowed further the effective administration of adaptive testing. Computerized Adaptive Testing (CAT) administers items tailored to the estimated ability of the examinee. The item level adaptation in CAT provides high measurement precision, while the computerized administration makes the tests more cost-effective (Weiss, 1982). Under CAT, each examinee takes a different set of items according to his or her ability. Since test developers cannot obtain in advance information regarding which items will be administered, CAT may lack quality control known as content balancing (Luecht & Nungester, 1998). The multistage test (MST) has drawn much attention and has succeeded in overcoming this issue.

Even in the early 1960s, the concept of MST emerged within the context of paper-and-pencil (P&P) format (Cronbach & Gleser, 1965; Linn, Rock, & Cleary, 1969; Lord, 1971). While the full CAT selects individual items for each examinee at a given moment, MST performs a group-of-items (i.e., modules) level adaptation. Modules are pre-assembled before test administration to satisfy both content balancing and the quality of the test structure. Panel, an independent unit of the test form in MST, is built in stages and there are two or three modules within each stage.

Compared to linear tests, MST produces a shorter test while still maintaining measurement accuracy (Betz & Weiss, 1974; Linn et al., 1969; Wainer, 1995) and also achieves a higher efficiency when a wide range of abilities needs to be measured (Lord, 1971, 1980; Patsula & Hambleton, 1999). In addition, since the tests are designed before their administration, MSTs can achieve strong control over the quality of the test and content balancing while maintaining a comparable measurement precision when compared to the results for CAT (Patsula & Hambleton, 1999). Unlike CAT, where examinees cannot modify answers for previous questions, MST allows test-takers to change responses to items within an individual module, because adaptivity of responses occurs at the module level, not at the item level.

MSTs have been researched and applied to purely dichotomously (e.g., true/false and multiple-choice) or polytomously (e.g., constructed-response, open-ended, performance-based, and essay) scored items. However, mixed-format tests are more frequently used. Mixed-format tests administer both dichotomously scored and polytomously scored items. Multiple-choice (MC) items are known to measure a broad range of domains, and scoring is easy and objective. On the other hand, constructed -response (CR) items tend to measure a relatively narrow range of domains while scoring is costly, time consuming and susceptible to more subjectivity than for MC items. Therefore, it might be beneficial to combine different types of items to strengthen the full psychometric properties of a test (Bennett, Morley, & Quardt, 2000; Breithaupt, Mills, & Melican, 2006; Hagge & Kolen, 2011; Zenisky & Sireci, 2002).

Even though the advantages of MST and mixed-format tests have been recently emphasized, there is as yet little empirical research on the benefits of using MST for mixed-format items. In addition, the application of MST to mixed-format tests raises other various issues that are worthwhile to consider. Researchers need to determine all the design aspects of MST within the context of a mixed-format item pool. For example, the test designer is required to determine the optimal test length, given measurement precision, the ratio between the different item types, the arrangement of chosen items in individual modules, and the method of scoring. This study seeks to evaluate the performance of mixed-format MST and answer important questions regarding the impact of the length of the test, module construction, and the impact of different proportions for the different item types on the measurement precision of the test.

Chapter 2

LITERATURE REVIEW

This literature review provides background information relevant to the current study. The first section explores the assumptions and characteristics of item response theory. The second section discusses the details of the multistage test (MST) and this includes the structural components of the MST, methods for its construction, and the algorithms used for for the MST administration. The third section summarizes the previous research pertaining to the mixed-format test, which is the target test for MST application in this study. The final section provides the statement of the research problem.

Item Response Theory

Item response theory (IRT) describes the probabilities of response outcomes through examining a set of item parameters and the examinee's trait level (Lord, 1952). Thus, IRT relates examinees and test items through mathematical models (Wainer & Mislevy, 1990). Item parameters may include difficulty and discrimination parameters, where item difficulty parameter describes the relative difficulty or easiness of the item, and an item discrimination parameter which indicates how well an item distinguishes between examinees with different levels of proficiency. The trait level, in an educational setting, is realized as a latent variable that represents an individual's ability level within a specified domain of knowledge For a medical measurement, on the other hand, the latent variable may indicate the level of physical or psychological traits of examines, such as pain or anxiety.

IRT models are commonly categorized into two classes based on the way that item responses are scored. Dichotomous IRT models have only two response options. Multiple-choice (MC) items and true-false items are modeled as dichotomous IRT models. Items with more than two response categories are modeled as polytomous IRT models. Polytomous IRT models are appropriate when responses are allowed to receive partial credit for partially correct solutions. Therefore, the response categories are ordered so as to represent the amount of the latent trait being measured.

IRT Assumptions

In order for an IRT model to be applicable for valid inferences, three assumptions should be satisfied: dimensionality, local independence, and functional form assumption (Embretson & Reise, 2000).

The first assumption, namely that of dimensionality, states that the IRT model contains the correct number of latent traits per examinee. When a single latent trait explain the examinee' response to items, the model is called unidimensional IRT model (Embretson & Reise, 2000). Very few tests contain items that are strictly unidimensional, but these tests are a fairly good approximation to unidimensionality (Lord, 1963). Factor analytic techniques are often used to test this assumption.

The second assumption is local independence (Hambleton & Swaminathan, 1984). It posits that the statistical dependence among items is explained away by the parameters of the model parameters conditional on the trait level (Lord & Novick, 1968). Under the local independence assumption, the probability of responding to one item is statistically independent of the probability of responding to any other item for a given ability level. Therefore, the probability of an examinee's response pattern equals the product of the probability of a given response to each item (Hambleton & Swaminathan, 1984). Through local dependence, any subset of items that measure the same trait can be used to estimate examinees' ability levels. This is one of the key IRT properties that enables the adaptive nature of CAT or MST. Local independence is violated when the content provided in one item gives a cue or information to the answer of another item in the same set of items.

The third assumption, functional form, is concerned with the item characteristic curve (ICC). The ICC is the mathematical function relating the probability of success for an item to the ability it measures (Hambleton & Cook, 1977). Different IRT models produce different ICCs based on their model parameters and the mathematical forms (Embretson & Reise, 2000).

Dichotomous IRT Models

In educational assessments, dichotomously scored test items (e.g., correct/incorrect, or true/false) are commonly used. As the dominant form of these scored items, MC items offer numerous merits, such as wide content coverage, high reliability achieved through a sufficient number of high-quality test items, and an ease of administration and scoring (Haladyna, 2004). Dichotomous IRT models are commonly applied to items whose responses are classified into binary categories, such as a MC item. The three most common dichotomous IRT models are: 1) the one-parameter logistic (Rasch, 1960); 2) the two-parameter logistic (Birnbaum, 1968); and 3) the three-parameter logistic (Birnbaum, 1968).

The one-parameter logistic model (Rasch, 1960) is the most parsimonious of the IRT models. This model assumes there is the unit discrimination for all items and guessing does not exist. The probability of passing item i by examinee j using the one-parameter logistic model is

$$P_{ij}(x_{ij} = 1|\theta_j) = \frac{1}{1 + exp(-(\theta_j - b_i))}, \tag{2.1}$$

where b_i is the item difficulty parameter and θ_j is the latent variable for the person's ability. The probability of passing the item increases as the value of θ_j increases for a given b_i, thus forming an S-shaped function along the latent trait scale. The one-parameter logistic model is also known as the Rasch model (Rasch, 1960).

The two-parameter logistic IRT model has difficulty (b) and discrimination (a) parameters. The probability of success (x=1) for a person j with a given ability (θ_j) on item i for the two-parameter logistic model is

$$P_{ij}(x_{ij} = 1|\theta_j) = \frac{1}{1 + exp(-a_i(\theta_j - b_i))}, \tag{2.2}$$

where a_i is the discrimination parameter, b_i is the item difficulty parameter for item i respectively, and θ_j is the variable for the person j's ability. The two-parameter logistic model assumes that guessing does not contribute to the item response.

The three-parameter logistic IRT model is the most general form of the dichotomous IRT models, and it has three parameters: difficulty (b), discrimination (a), and the pseudo-guessing parameter (c). The probability of success (x=1) for a person j with a given ability (θ_j) on item i in the three-parameter logistic model is

$$P_{ij}(x_{ij} = 1|\theta_j) = c_i + (1 - c_i)\frac{1}{1 + exp(-a_i(\theta_j - b_i))}, \tag{2.3}$$

where a_i is the discrimination parameter, b_i is the item difficulty parameter, c_i is the pseudo-guessing parameter for item i respectively, and θ_j is the variable for the person j's ability.

For dichotomous IRT models, the probability of passing an item increases monotonically as the value of θ increases for given item parameters. This S-shaped function along the θ is known as the item characteristic curve (ICC). The ICC offers a graphical representation of the probability of a correct response for each item along the latent trait scale (Embretson &

Reise, 2000). Figure 2.1 shows the ICC for the three-parameter logistic model where a=1.5, b=0.5 and c=0.1.

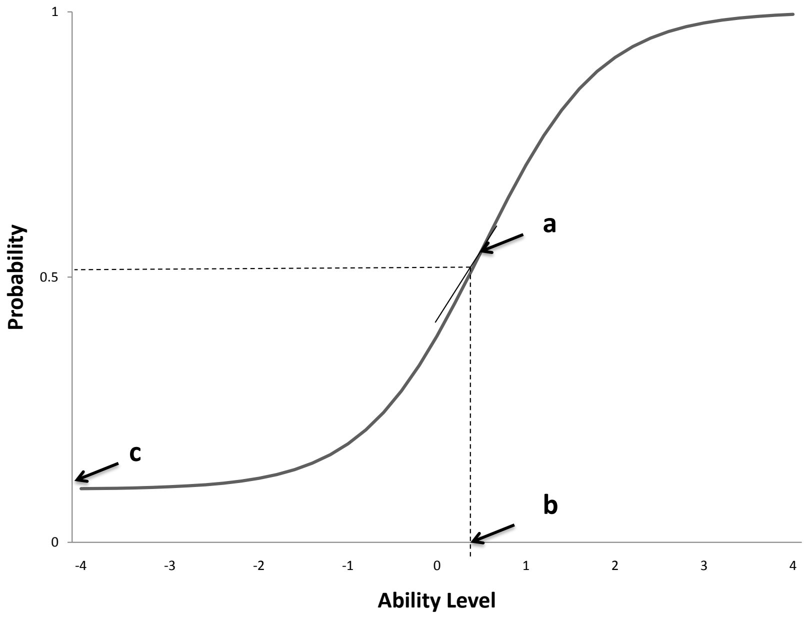

Figure 2.1: Item Characteristic Curve for the Three-Parameter Logistic Model

The difficulty parameter b indicates the relative easiness of an item, typically ranging from -3 to +3 on the ability scale. Positively large difficulty parameter values indicates that the items are relatively difficult. The point at which the slope of the ICC is maximal is known as the point of inflection, and the difficulty parameter for the one-, two-, and three-parameter logistic IRT models is defined as the θ value that corresponds to the point of inflection. The probability of answering the item correctly at the point of inflection is $\frac{1+c}{2}$ for the three-parameter logistic model.

The discrimination parameter a indicates how well an item distinguishes examinees with high proficiency from examinees with low proficiency. The discrimination parameter is proportional to the slope of ICC at its point of inflection. Therefore, a large discrimination parameter value indicates a steep slope for ICC. A steep slope for ICC implies that the item is able to distinguish between ability levels in the vicinity of the difficulty parameter (Hambleton & Swaminathan, 1984). Typical a values range between 0 and around +2, even though theoretically these values can be infinitely large.

Since examinees are asked to choose one of the options presented in MC items, low ability examinees may guess. Under the three-parameter logistic model, the c parameter reflects item-dependent pseudo-guessing. A high c value indicates a high probability of getting an item correct for individuals with low ability. One interpretation of the three-parameter logistic model is that item responses are composed of two processes: a p-process and a g-process (Hutchinson, 1991). The p-process is an item-solving process whereas the g-process is a guessing process. One possible arrangement for executing the two processes is that the g-process is followed by the p-process: As subject attempts to solve an item and resorts to guessing, only if a correct response is not identified. With ICC, c is the lower asymptote as shown in Figure 2.1.

Item and Test Information for the Dichotomous IRT Models

IRT provides a precision of measurement for individual items across the latent trait levels through the use of an item information function. The information function of dichotomously scored items under IRT is expressed as:

$$I(\theta) = \frac{P'(\theta)^2}{P(\theta)(1 - P(\theta))} \tag{2.4}$$

, where $P(\theta)$ is the probability of passing the item given θ, and $P'(\theta)$ is its first derivative. For the three-parameter logistic model, Equation 2.4 can be expressed as:

$$I_{3PL}(\theta) = a^2 \frac{Q(\theta)}{P(\theta)} \left[\frac{P(\theta) - c}{1 - c} \right]^2, \tag{2.5}$$

where $Q(\theta) = 1 - P(\theta)$. From the Equation (2.5), the 2PL information function can be calculated by setting c=0 while the 1PL information function is obtained by setting c=0 and a=1 simultaneously. The information function for the dichotomous IRT models is typically bell-shaped curve with positive values. For 2PL, the information is largest at the item difficulty while the amount of information decreases as the θ moves toward either extreme.

The information function for a test (i.e., called test information function) can be expressed as the sum of the individual information function of each item conditional on θ.

$$TI(\theta) = \sum_{i}^{I} I_i(\theta), \tag{2.6}$$

where $I_i(\theta)$ is the information function for item i and I is the item number. The standard error of measurement for a test conditional on θ is expressed thus using the test information function, $TI(\theta)$, as:

$$SE(\theta) = \frac{1}{\sqrt{TI(\theta)}}. \tag{2.7}$$

Large information at θ indicates that the item is able to measure the examinee's ability (θ) with high precision. For the one-parameter logistic IRT model, the items have the same amount of item information (i.e., 0.25) at the "peak" of the information function (i.e., where $\theta = b$). However, for the two-parameter logistic IRT models, the information function is the function of the item discrimination parameter. The item information functions for dichotomous logistic IRT models are smooth and continuous along the ability scale, while the peaks of one- and two-parameter logistic IRT models form around the difficulty parameter values (van ver Linden & Boekkooi-Timminga, 1989). For three-parameter logistic IRT model, the location of the peak shifts above item difficulty as the value of the pseudo-guessing parameter increases (Lord & Novick, 1968). Figure 2.2 illustrates the information function for the three-parameter logistic model where a=1.5, b=0.5 and c=0.1.

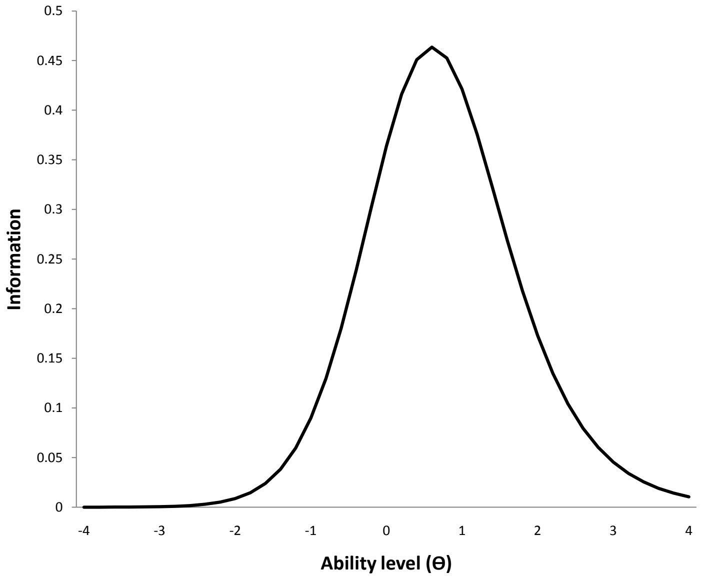

Figure 2.2: Item Information Function for the Three-Parameter Logistic Model

Polytomous IRT Models

While dichotomous IRT models are used for binary response categories, polytomous IRT models allow multiple ordered response categories. In this way, polytomous IRT models can distinguish between examinees who have complete knowledge and those who have incomplete, but a certain amount of knowledge. For example, an item with partial credit can award points for an answer that is not eligible for the full credit and an essay can grant more points for better answers according to a scoring rubric. Since polytomous items have a larger number of response categories, they offer more information along a broader range of ability scale (Ostini & Nering, 2006).

According to Thissen and Steinberg (1986), polytomous IRT models can be classified as two types: difference models and divide-by-total models. Applying this categorization, the graded response model and its variations are examples of the difference models, and the probability of answering a category score is then determined by calculating the difference between two adjacent probability curves called category characteristic curves (Dodd, 1984) . The partial credit model and the generalized partial credit model are categorized as divide-by-total models, because the probability of obtaining a category score is divided by the sum of the probabilities for all possible category scores for that item (Dodd, De Ayala, & Koch, 1995; Thissen & Steinberg, 1986). A few representative models that are commonly used are described in the following sections.

Graded Response Model

The graded response model (GRM; Samejima, 1969) is used for items when two or more response categories are so ordered to indicate the examinee's level of proficiency. Responses x to item i in this model are scored from 0 to m_i, resulting in $m_i + 1$ score categories. Lower score values reflect less of an examinee's proficiency and higher values reflect more proficiency in the domain that under the item is measuring. Development of the GRM model is expressed as two steps: boundary response probabilities and category response probabilities. The boundary response probability is defined as the probability of an examinee with a given ability to respond with a category score x or higher for an item. The boundary response probability for a response of x or higher for item i is expressed as :

$$P^*_{ix}(\theta) = \frac{1}{1 + exp(-a_i(\theta - b_{ix})}, \tag{2.8}$$

where a_i is the discrimination parameter for item i and b_{ix} is the category boundary for category score x for item i on the θ scale.

In the second step, the category response probabilities are formed to determine the probability of an examinee responding with a category score x for an item. Mathematically, the category probability for the category x is the difference between two adjacent boundary response probabilities at x and $x + 1$. The category probability for response x for item i is

$$P_{ix}(\theta) = P^*_{ix}(\theta) - P^*_{i(x+1)}(\theta), \tag{2.9}$$

where $P^*_{ix}(\theta)$ is the probability of responding in x or higher for item i. The boundary probability of responding to the lowest category or higher (i.e., the entire category) is 1.0 by definition, while the probability of responding above the highest category is 0.0. Figure 2.3 shows the category response probabilities for an example item with four ordered responses (i.e., three category boundaries, b_1=-1.0, b_2=0.0, and b_3=1.0) with a=1.5.

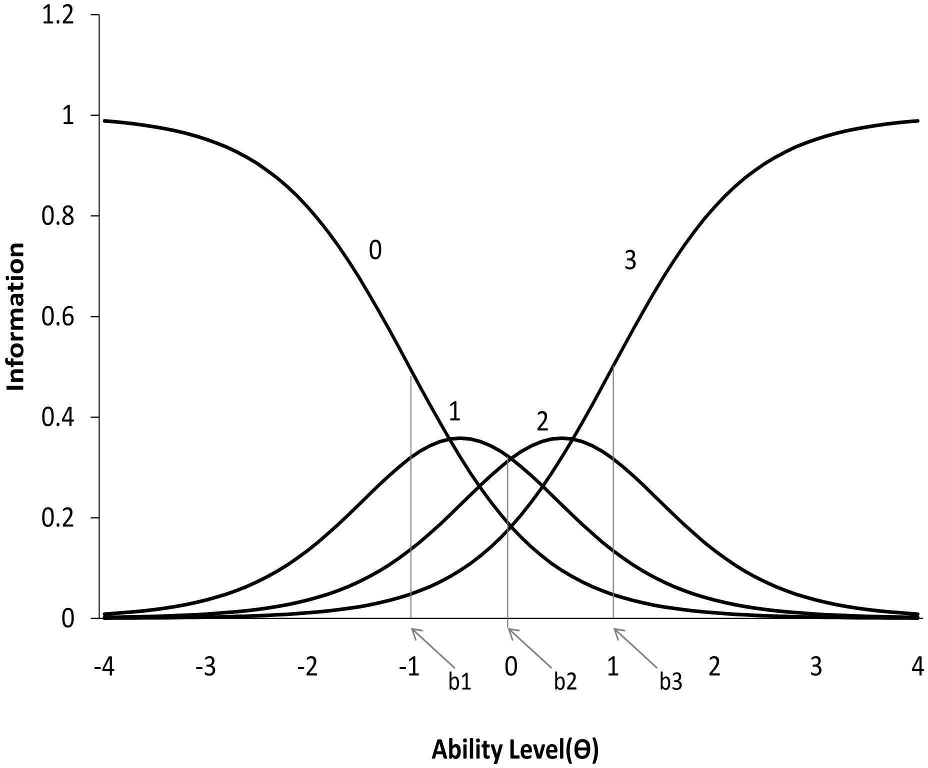

Figure 2.3: Category Response Probability for the Graded Response Model

Partial Credit Model

The partial credit model (PCM; Masters, 1982) was developed to analyze the item types in which examinees are required to complete multiple steps during the solution process. Those steps are called step difficulties. In an item with four steps, for instance, there are three step difficulties. Step difficulties indicate how difficult it is to transition from one response category to the next and can also be referred to as category intersections (Embretson & Reise, 2000). The partial credit model assumes that examinees must complete these steps in a sequential order. For example, an examinee cannot achieve the credit for step 3 without successfully completing steps 1 and 2. While category boundaries in GRM are required to be in order on the θ scale, step difficulties in the partial credit model do not have to be so ordered. For instance, step 2 may be more difficult to complete than step 3. A "reversal" occurs when the difficulty levels of the steps are not in order (Dodd & Koch, 1987).

As an extension of the Rasch model to the polytomous case, the PCM assumes all items have equal discrimination. The probability of an examinee obtaining a category score x on item i is thus defined as:

$$P_{ix}(\theta) = \frac{exp(\sum_{k=0}^{x}(\theta - b_{ik}))}{\sum_{r=0}^{m_i}(exp(\sum_{k=0}^{r}(\theta - b_{ik})))}, \tag{2.10}$$

where b_{ik} is the step difficulty of score category k for item i, and m_i is the number of score categories minus one for item i.

Generalized Partial Credit Model

The generalized partial credit model (GPCM; Muraki, 1992) generalizes PCM by allowing items to have different discrimination power. The probability of an examinee with a given ability responding with a category score x for item i is defined as :

$$P_{ix}(\theta) = \frac{exp(\sum_{k=0}^{x} a_i(\theta - b_{ik}))}{\sum_{r=0}^{m_i}(exp(\sum_{k=0}^{r} a_i(\theta - b_{ik})))}, \tag{2.11}$$

where b_{ik} is the step difficulty of score category k for item i, m_i is the number of score categories minus one for item i, and a_i is the discrimination parameter for item i. Similar to GRM, the GPCM has a single discrimination power for an item. As PCM is an extension of the Rasch model, GPCM is an extension of the two-parameter logistic model. When there are two response categories, GPMC is reduced to the two-parameter logistic model (Muraki, 1992).

Item Information for Polytomous IRT Models

The polytomous IRT models deliver greater precision of estimated ability by taking advantage of the information in each response category of an item (Hambleton & Swaminathan, 1984). Since polytomously scored items involve multiple category responses, each score category contributes to the item information function (Dodd et al., 1995). According to Samejima (1969), the item information function of a polytomous IRT model is the sum of the category information functions, defined as the information contributed by a category to an item. This category information function for item i

category x is expressed as:

$$I_{ix}(\theta) = \frac{P'_{ix}(\theta)^2}{P_{ix}(\theta)^2} - \frac{P''_{ix}(\theta)}{P_{ix}(\theta)}, \tag{2.12}$$

where $P_{ix}(\theta)$ is the probability of responding with a score category x for item i at θ, and $P'_{ix}(\theta)$ and $P''_{ix}(\theta)$ are the first and second derivatives of $P_{ix}(\theta)$ respectively. Item information, then is obtained by adding the category information functions weighted based on the probability of each score category (Dodd et al., 1995; Samejima, 1969), written as:

$$I_i(\theta) = \sum_{x=0}^{m_i} \frac{[P'_{ix}(\theta)]^2}{P_{ix}(\theta)}, \tag{2.13}$$

where m_i is the number of category scores for item i.

For polytomous IRT models, the locations of the step difficulty or category boundaries determine how far the information function spreads across θ. A large distance between b_{ik}s results in the information function spanning a wider range on the θ scale (Dodd & Koch, 1987; Koch, 1983). In addition, compared to the dichotomous items, polytomous items in general provide more information along the θ continuum (Embretson & Reise, 2000; Jodoin, 2003a; Muraki, 1993; Ostini & Nering, 2006).

The test information function, as was the case with the dichotomous IRT models, is defined as the sum of the item information conditional on θ (see Equation 2.6). Similarly, the measurement precision of a test is evaluated using the standard error of measurement, which is expressed as the square root of the reciprocal of the test information (see Equation 2.7).

Multistage Testing

One of the main challenges for mass-administered testing is that examinees will exhibit a broad range of ability. A set of items within a linear test may be too easy for highly proficient examinees or too difficult for examinees of lower ability. One way to overcome this mismatch between test and examinees' abilities is to administer items that are the most informative for individual examinees abilities. Computerized adaptive testing (CAT) tailors a test for each examinee by estimating that examinee's ability based on responses to previous items and then choosing an item that will increase measurement accuracy for the test. Since each examinee answers only items highly relevant to his or her ability level, CAT can achieve test length reduction without sacrificing measurement precision (Bergstrom & Lunz, 1999).

Unlike CAT, multistage testing (MST) administers a set of items (i.e., a module) to the examinees at each adaptation point. Thus, adaptation occurs at a module level, not at the level of individual items. One of the advantages of MST over CAT is that it offers the possibility of a strong quality control for items on the test form. While the test form for CAT is generated during the test administration, MST test forms are constructed before such administration. Therefore, a strong quality control for items on the test forms can be achieved through evaluation by test experts and content committee experts (Hendrickson, 2007; Luecht & Nungester, 1998; Patsula, 1999). The following sections describe the details of MST including its components and panel structure, followed by a discussion of the ability estimation and routing methods. In addition, MST construction technique relevant to the current study is presented.

MST Components

Panels, stages, modules, and pathways are often considered as building blocks of MST (Luecht, 2000). Panels are top level components, which are essentially a test form to which an examinee is assigned. Each panel should be equivalent in terms of its information and content characteristics, so that tests are fairly administered to all examinees. Within panels, there are commonly two or three stages. A stage is a collection of modules, and an examinee must finish a module in one stage in order to advance to a subsequent module in the next stage. Modules within a stage are constructed to represent distinct difficulty levels so the adaptive administration of a set of items can be accomplished. Modules can be specified as easy, medium and hard modules based on their overall difficulties, which is determined from the average item difficulty within the individual module. The first stage commonly has a single medium difficulty module, while the second and third stage consists of two or three modules.

Examinees are administered one of the panels in MST. After finishing a module in each stage, each examinee follows a pathway to continue onto the next stage. Pathways are the allowed sequences of modules in a test. Pathways allow modules to be administered to different examinees according to their ability and the routing rules. Figure 2.4 illustrates a MST panel with seven pathways and each examinee is assigned to these pathways during administration of test. For example, high ability examinees are more likely to be routed to pathways that sequence through hard modules (e.g., Pathways 6 or 7), which will provide more information for examinees of high abilities. A routing method refers to a systematic method for choosing pathways for each examinee. It is this routing method through which the adaptation of MST is implemented.

MST Panel Structure

Figure 2.4 illustrates an example of MST panel structure in which there are a total of seven modules and seven pathways. It has one module at the first stage (i.e., routing module) and three modules each at the second and third stages, while modules at the second and third stages represent easy, medium and hard difficulty. This panel structure is known as 1-3-3 MST and is widely employed in both research and practice (Davis & Dodd, 2003; Hambleton & Xing, 2006; Jodoin, Zenisky, & Hambleton, 2006; Luecht, Brumfield, & Breithaupt, 2006; Luecht & Nungester, 1998).

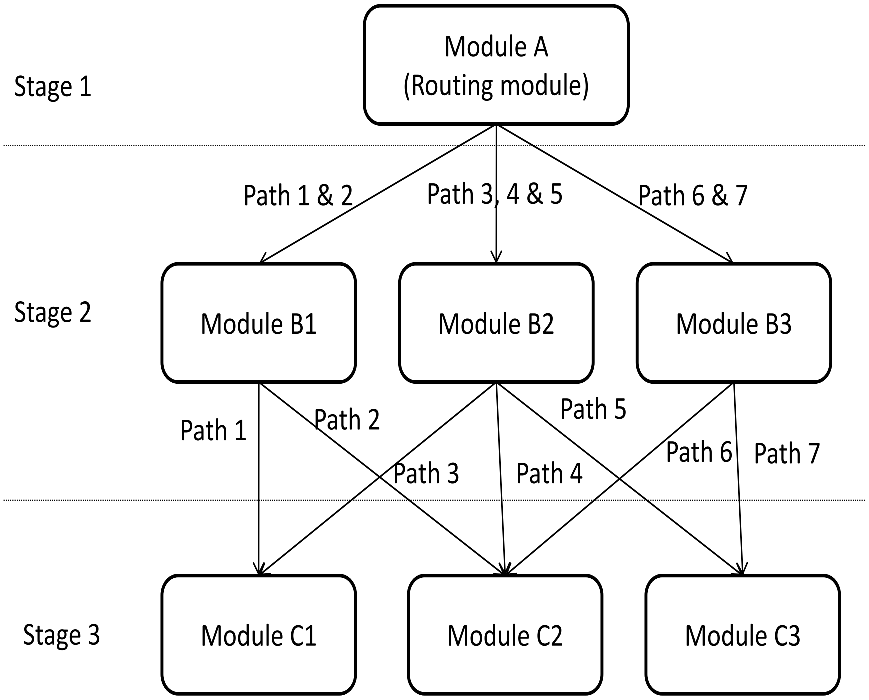

Figure 2.4: The 1-3-3 Panel Structure for MST

Pathways, the potential sequences of modules that an examinee may take across different stages, is another element of MST panel structure. The first stage is commonly known as a routing test, and it typically contains one module. All examinees assigned to a panel will receive this module. There are often two or three modules for the second and third stages to cover necessary difficulties and meet the purpose of the test. A test developed for ability estimation often has three modules for the second and third stages (i.e., 1-3-3) and covers a wide range on the θ scale.

MST Ability Estimation

After all the items in a module are administered, interim or *provisional* estimates of θ are needed to choose the next module. In the context of an IRT-based ability estimation, two types of ability estimation methods are commonly used–maximum likelihood estimation (MLE) and Bayesian estimation, such as expected *a posteriori* (EAP) or maximum *a posteriori* (MAP).

Maximum likelihood estimation (MLE) searches for a point estimation of the ability that maximizes the likelihood of the response pattern. Given the known item parameters, the ability estimate that maximizes the likelihood of the response pattern for each examinee is calculated. A fundamental assumption for this approach is conditional independence, which states that the probability of responding to an item correctly is independent of responding to another item, thus holding the ability (θ) constant. When this assumption is violated, MLE tends to overestimates the accuracy of such estimation (Thissen, Steinberg, & Mooney, 1989).

Maximum a posterior (MAP; Samejima, 1969) and expected a posterior (EAP; Bock & Mislevy, 1982) are two popular methods that belong to the Bayesian ability estimations. The Bayes theorem states that the posterior distribution of ability is proportional to the product of the likelihood function and the prior. The posterior distribution of ability according to Bayes' theorem is:

$$P(\theta|Y_j) = \frac{P(Y_j|\theta)g(\theta)}{\int P(Y_j|\theta)g(\theta)d\theta} = \frac{P(Y_j|\theta)g(\theta)}{P(Y_j)}, \tag{2.14}$$

where Y_j is the response string for the jth examinee for items in the test and $g(\theta)$ is the prior distribution for θ.

MAP calculates the mode of the posterior through a numerical method, such as the Newton-Raphson method. The measurement precision of the MAP ability estimate generally exceeds that of MLE because the prior augments the information on the ability parameter (Wainer & Mislevy, 1990). Since the posterior is the product of likelihood and prior, MLE can be considered a special case of MAP with a uniform prior. Mathematically, the MAP estimates can be obtained by solving the derivative of the log-posterior distribution with respect to θ, which is expressed as:

$$\frac{\partial log[P(\theta|Y_j)]}{\partial\theta} = \frac{\partial log[P(Y_j|\theta)]}{\partial\theta} + \frac{\partial log[g(\theta)]}{\partial\theta} \tag{2.15}$$

where Y_j is the response string for the jth examinee for items in the test and $g(\theta)$ is the prior distribution for θ.

EAP is another Bayesian estimation method that calculates the mean for the posterior distribution of ability (Bock & Mislevy, 1982). The measurement precision for the ability estimate is calculated from the standard deviation of the posterior distribution of θ. The mean and standard deviation are approximated through numerical methods (i.e., the weighted summation for the values of the function on quadrature points). A numerical approximation is performed by taking the n quadrature points along the ability scale. The weights are typically drawn from the approximation of a known distribution, such as the standard normal distribution (Embretson & Reise, 2000). Unlike MLE or MAP, EAP a is non-iterative method, and the accuracy of the estimation can be improved by increasing the number of quadrature points. An EAP estimate can be obtained from the following formula:

$$\hat{\theta} = \frac{\sum_{r=1}^{q} [Q_r \times L(Q_r) \times W(Q_r)]}{\sum_{r=1}^{q} [L(Q_r) \times W(Q_r)]} \tag{2.16}$$

where Q_r is rth quadrature node, $W(Q_r)$ is the weights at quadrature node Q_r serving as a discrete prior distribution, and $L(Q_r)$ is the likelihood function evaluated at Q_r.

Unlike MLE, which cannot find a solution for θ when an examinee answers all items correctly or incorrectly, MAP and EAP always find a solution. In addition, MAP and EAP are more stable for short tests (Kingsbury & Weiss, 1983). However, the Bayesian methods tend to underestimate high abilities and overestimate low abilities (Parshall, Spray, Kalohn, & Davey, 2002). In addition, a correct prior specification is important, and a wrong prior may result in biased estimates (Wainer & Thissen, 1987).

MST Routing

The routing method governs which module in the next stage will be administered for an examinee. Through routing, test designers may succeed in achieve administering the most informative modules delivery to for each examinee.

A defined population interval (DPI) method (Luecht et al., 2006) routes an examinee according to the predefined population proportions using an estimate of his or her provisional ability. Assuming that equal proportions of examinees are expected to be routed to three major pathways for 1-3-3 MST, then two estimated true-scores (i.e., X_1 and X_2) are obtained from the following equations:

$$X_1 = \sum_{i \in 1M} P(\theta_1; \xi_i) \tag{2.17}$$

$$X_2 = \sum_{i \in 1M} P(\theta_2; \xi_i), \tag{2.18}$$

where $1M$ is the set of items in the routing module, ξ_i is the item parameters for item i. For the normally distributed population ability, -0.44 and 0.44 may be used for θ_1 and θ_2 respectively. Finally, an examinee's number-correct score is compared to the estimated true-scores (i.e., X_1 and X_2) to determine which routing path that individual examinee takes. This method has been widely implemented in previous studies (Jodoin, 2003b; Xing, 2001; Zenisky, 2004). The main advantagee of DPI are the control available for item exposure and MST path utilization.

The approximate maximum information (AMI) method (Luecht et al., 2006) performs the routing based on the intersections of cumulative test information functions (TIFs) between neighboring modules within a stage. For examinees who finished the first stage of the 1-2-2 MST design, the intersection between the accumulated information function of the medium-medium pathway (i.e., medium modules at the first and the second stage) and the medium-hard pathway (i.e., a medium module at the first stage and a hard module at the second stage) is calculated. The intersections on the ability scale are translated to an estimated true-score, which will eventually be compared to the examinee's number-correct score for the routing decision.

The item selection method under CAT can also be used for MST routing; the module providing the largest amount of information at the current ability estimate is then chosen (Davis & Dodd, 2003). This method, called a modified AMI (M-AMI), is optimal in terms of the IRT information function, but may suffer from disproportional exposure rates for different modules (Kim, Chung, Park, & Dodd, 2013). For example, if examinee abilities are normally distributed, then the exposure rate for medium-difficulty modules will likely be greater than for the easy- or hard-difficulty modules.

MST Construction

Multiple panels are often required to meet content specifications and control item exposure rates for test security. Manual construction of multiple panels often takes a prohibitively long time even with a small number of panels, and it is almost impossible for any panels of practical numbers (Luecht et al., 2006; Luecht & Nungester, 1998). Thus, MST multiple panel constructions are usually performed using computer algorithms commonly referred as automated test assembly (ATA) techniques.

ATA selects a set of items for the modules to build panels according to panel structure and according to various constraints, including test length, test information, content constraints and exposure rate controls. Content specification distributes item content area in the panel and provides content validity evidence according to a test blueprint. Exposure rate control is needed because items should not be administered above a predefined frequency of utilization. Over-exposed items threaten item security and compromise the validity of the test (Revuelta & Ponsoda, 1998). Target TIF is the main method used to specify the amount of test information in a MST design, which is directly related to the measurement accuracy of the test. For high measurement precision, a large TIF for a panel is desirable. However, simultaneous construction of multiple panels requires test developers to consider methods that can deliver a fair allocation of item information among the panels. In order to assess a practical TIF that can be supported by the item pool, test developers often perform preliminary panel constructions before actual test construction (Luecht & Burgin, 2003; Zenisky, 2004). Two of the most popular ATA algorithms are linear programming (LP) and heuristics. LP relies on a mixed-integer linear programming solver that searches for a set of items that meet the constraints while an objective function is optimized. Heuristics, on the other hand, iteratively choose items to form a test that satisfies the statistical and other constraints of the test. The current study will assemble MST based on the LP, and so the details of the LP method are presented.

LP is the optimization problem of an objective function under multiple constraints. The unique characteristic of LP problems is that objective and constraints are expressed in linear equalities or inequalities of the decision variables (DVs). The goal of the LP solver is to find the values of decision variables (DVs) that optimize the objective function while yet satisfying all constraints. For example, an LP problem may express the profit that needs to be maximized under the resource constraints. DVs are real values if they represent the unit of resources for which fractions are allowed (e.g., the amount of dollars, the weight of fuel), while integer DVs are used when they represent objects where their fractions do not have any physical meaning (e.g., the number of workers, the number of items on the test). LP is called feasible if optimization is achievable while satisfying all constraints. The test assembly problem can be expressed as an LP problem in which DVs represent binary variables that indicate whether the corresponding items are selected or not selected (Theunissen, 1986). A DV of one indicates the item is chosen for the test form.

Under the LP framework, TIF provides a reference point for the test information along the ability scale (Boekkooi-Timminga, 1990). The objective function in the LP model often involves specifying optimality requirements for tests in terms of the TIFs. Two types of a target TIF are commonly programmed; a relative target TIF and an absolute target TIF (van der Linden, 2005). A relative target TIF does not provide absolute information values for the assembled test. Instead, it defines a set of numbers that represent the amount of information relative to other points. One of the benefits of using the relative target TIF method is that the test designer does not have to be familiar with the scale of the information function. Instead, the overall shape of the test information function becomes the statistical constraint for the test construction. Therefore, the relative target TIF is preferred when the test designer wants to construct a test with a specific shape and maximum information for high measurement precision. The shape can be uniform for a norm-referenced test or peaked for a criterion-referenced test. An absolute target TIF, on the other hand, specifies the exact height of information across the latent trait scale. When parallelism among tests is required, the absolute target TIF is often used to control the amount of information among tests. This LP model minimizes the maximum absolute deviation between an absolute target TIF and the assembled test.

In addition, LP models are known to provide a flexible presentation of numerous constraints (van der Linden, 1998, 2005). For instance, the number of items on the test can be expressed as an equality:

$$\sum_i x_i = n, \tag{2.19}$$

where x_i denotes the DV for item i, and n is the number of items in the test. Since x_i is only one if item i is chosen for the test, then the the summation of x_i signifies the number of items on the test.

Similarly, if n_c items should be chosen from the content c, and then the following constraint can be added :

$$\sum_{i \in V_c} x_i = n_c, \tag{2.20}$$

where V_c is the item set for content c, and n_c indicates the target number of items of content c in the test.

If V_e denotes the set of items required to be excluded from the test construction, then the following constraint is added :

$$\sum_{i \in V_e} x_i = 0. \tag{2.21}$$

The mutual exclusive rule implies that if one item is chosen, then the other should not be included. The corresponding LP model that prevents item i and j from being included on the test simultaneously is:

$$x_i + x_j < 2. \tag{2.22}$$

The absolute target TIF specifies the exact height of information across the latent trait scale. Because the content areas and the TIFs determine the parallelism among tests (Lord, 1977; Samejima, 1977) in the IRT environment, the absolute target TIF is used when parallel tests are assembled. This modeling minimizes the maximum absolute deviation between an absolute target TIF and the assembled test. The current study will perform MST construction using absolute target TIFs.

Mixed Format Tests

Many assessments in practice include both MC items and constructed-response (CR) items and thus, they are called mixed-format tests. Examples of the relatively well-known mixed-format tests are Advanced Placement (AP) subject examinations, National Assessment of Educational Progress (NAEP) exams, and assessment tests in a few states, such as North Carolina and Wisconsin (Reshetar & Melican, 2010; Rosa, Swygert, Nelson, & Thissen, 2001). Mixed-format tests benefit from advantages of both MC items and CR items. Tests that consist of MC items are known to cover a broad range of content while being relatively inexpensive to score. On the other hand, tests with CR items tend to focus on a relatively narrow range of content while they are able to measure complex skills and a higher level of performance (Hogan & Murphy, 2007). A more balanced test can be achieved by combining two different item formats (Bennett et al., 2000; Breithaupt et al., 2006; Ercikan et al., 1998; Wainer & Thissen, 1993; Zenisky & Sireci, 2002).

Previous Studies

A large volume of research studies have been conducted for various design aspects of MST. This section presents previous research studies and research questions of the current study. Previous research is organized by the factors impacting the measurement precision of MST, such as test length, MST panel structures, item pool, and MST construction methods.

Test Length

Among others, test length is one of the important factors that impact MST test performance. Many studies based on dichotomously-scored items have investigated the test length issue such as the number of items within the module and the interaction between total test length and the number of stages (e.g., Jodoin, 2003b; Jodoin et al., 2006; Luecht & Nungester, 1998; Patsula, 1999; Xing & Hambleton, 2004)). Results from these studies correspond to a well-known characteristic of testing: longer test lengths lead to an increase in the decision consistency or accuracy (Crocker & Algina, 1986). In terms of studies based on polytomously-scored items, Macken-Ruiz (2008) investigated the same, decreasing, or increasing number of items per stage in MST using items calibrated according to GPCM. Chen′s (2010) study, which is based on GPCM, investigated test lengths on the first stage (i.e., long routing and short routing) and reported that shorter routing test length tended to result in slightly better measurement precision. For studies using mixed-format items, Kim (2010) studied three test length variations. Specifically, the test length calculation of mixed-format tests on this study considered the percentage of each test unit type (i.e., dichotomous, three-category, and four-category test units) in the pool. As expected, the research showed a longer test length produced better results in terms of classification accuracy.

MST Panel Structure

As one of the crucial design aspects of MST, panel design arrangement is related to how stages and modules are constructed for the panel. Often, the number of stages and modules included in each panel depends on the test developers decision considering the testing purposes and policies. Zenisky, Hambleton, and Luecht (2010) argued that module/stage arrangements are influenced by the item pool capacity and features, the range of adaptivity based on estimating the examinees ability, and the precision of measurement ultimately desired from the test length and information function, among other factors. In addition, Hendrickson (2007) pointed out that the panel structure with more stages and modules provides better flexible adaption to each examinee. Using dichotomously-scored items calibrated according to the 3PL IRT model, Patsula (1999) compared various MST test designs and reported that the increasing the number of stages introduced higher accuracy in estimating ability. Furthermore, increasing the number of modules within the stage produced better ability estimates and efficiency. Zenisky (2004) studied variations on module arrangements, but different module arrangements did not produce significant differences in terms of the classification decision. Regarding studies based on polytomously-scored items, Chen (2010) examined the effects of eight MST test structures on ability estimation using the GPCM, and reported that all designs produced similar results in their ability estimation. In the mixed format tests, however, only the 1-3-3 (Kim, 2010; Kim & Dodd, 2010) and 1-3 (Oranje, Mazzeo, Xu, & Kulick, 2014) panel design have been investigated to date, and various panel designs have not yet been implemented in many MST research studies.

Item Pool

MST test performance has been reported to be impacted by the item pool design, because the item pool capabilities and capacities affect MST design elements such as test length; stage/module designs; shapes, heights, and locations of target test information functions. Jodoin's (2003b) studied the impact of item pool quality to the MST performance and found that enhanced item pool quality resulted in better measurement precision and classification accuracy. Others conducted studies manipulating the item pool by increasing the pool sizes or changing the characteristics of the item pool based on dichotomously-scored items (e.g., Jodoin, 2003b; Xing & Hambleton, 2004).

MST Construction Method

ATA procedures utilize the computer algorithms to construct MST panels. Common approaches for ATA include heuristic methods (Luecht, 2000), linear programming (van der Linden, 2005), and a network-flow procedure (van der Linden, 1998). Several recent studies have utilized heuristic-based procedure and linear programming for the MST designs using polytomously-scored or mixed-format items. Kim (2010) and Kim, Chung, Dodd, and Park (2012) implemented Luecht's (2000) normalized weighted absolute deviations heuristic (NWADH) and made a few modifications to meet the requirements of the research. In addition, Park, Kim, Chung, and Dodd (2011) used the LP solver to construct MST using the mixed-format test based on GPCM. A sequence for constructing modules was undertaken, and a fair distribution of items among panels was achieved by controlling the upper bounds in the branch-and-bound method (Land & Doig, 1960). Studies by Park, Kim, Chung, and Dodd (2011) and also Park, Kim, Chung, and Dodd (2012) used an ATA program called JPLEX (Park, Kim, Dodd, & Chung, 2011) to implement the actual MST construction using a mixed-format pool. As a part of the efforts to increase mixed-format item pool utilization, Park et al. (2012) proposed a new LP model. Multiple MST assemblies (i.e., MST reassembly) were performed by replacing a portion of the used items with unused items from the pool, thus increasing overall pool utilization. These results showed that the new method increased the overall pool utilization rates, while still meeting the expected statistical and non-statistical constraints.

Statement of Problem

A mixed-format MST test requires giving full consideration to various design components including test length, administration procedure, construction and scoring (Kinsey, 2003). In addition to those specific components, test administrators need to consider various proportional combinations of different item types based on the different purposes of each test for the mixed-format MST design. Accordingly, these different combinations of item types will impact test administration time, item utilization from the pool, and the scoring processes. To date, no study has investigated the impact of different proportions of item types on MST design.

These different proportions of item types in MST need investigation in terms of panel structure. The allocation of items can determine how these panel structures are constructed and also the influences that are apparent on the different aspects of MST design, such as measurement precision, test administration, and item pool utilization. To date, no study has investigated the various panel structures and the interaction between different proportions of item types and panel structures for a mixed-format MST. Further, the interaction between the proportions of item types and test lengths has never been studied. Finally, item type distributions at the routing module have not been considered in previous mixed-format MST studies. As these previous researchers (Jodoin et al., 2006; Kim et al., 2012; Zenisky, 2004) have pointed out, the routing test design may affect MST performance significantly. To fully utilize and better understand the nature of a mixed-format MST, various item type distributions at the routing module should be investigated further. Thus, there is a major need to evaluate and analyze the impacts of different proportions of item types for the mixed-format MST and also their interactions with other design elements, such as test length, panel structure, and item type distribution at the routing module to achieve better practicality and further psychometric enhancement of overall test administration.

Research Questions

The following research questions will be answered in this research study:

1. How does test length (total points) impact measurement accuracy for mixed-format MST?

2. What are the important design features when creating modules, stages, and panels for mixed-format MST?

3. How do the various MST structures under the mixed-format context differ in terms of measurement precision?

4. How do the varied proportions for item types in the mixed-format MST differ in the accuracy of ability estimation?

5. How does item type distribution at the routing module impact the measurement precision of mixed-format MST?

Chapter 3

METHODOLOGY

Design Overview

The term ”test unit” was used to describe both dichotomously scored and polytomously scored items. Test length, a common design factor for the previous MST research efforts (e.g. Jodoin et al., 2006; Kim, 2010; Zenisky, 2004), was indirectly controlled in this study through the condition of the total points. The current test unit pool contains test units with one-, two-, or three-step difficulties. These three test unit types contribute three maximum points (i.e., one, two, and three points) respectively to the total points of MST.

This study considered two levels of total points (i.e., 40 and 60). A test length of 60 dichotomous test units has often been recommended for the diagnostic and licensure tests (Jiao, 2003). The condition of 40 total points is included herein to approximately correspond to the test length used in Ho and Dodd's (2008) study, that was based on the high-stake mixed-format test.

Further, four different proportions of polytomous test unit scores in the MST were considered in this study. This condition includes MST with 10, 30, 50, and 70 percent of total points accounted for by polytomous test units to control the degree of contribution of these test units to the total points. In the current study, the dichotomous test units are scored either as 0 or 1, while the polytomous test units are scored from 0 to m_i (i.e., m_i+1 score categories).

In terms of MST panel structure, two conditions (i.e., 1-2-2 and 1-3-3) were considered for this study. The 1-2-2 MST structure is popular for classification testings, while 1-3-3 MST forms are most commonly researched for ability estimation testings (Jodoin et al., 2006; Zenisky, 2004).

Finally, two approaches for test unit type allocation for the routing module (i.e., module at the first stage) were considered in this study. The routing module was constructed by using either the mixture of dichotomous and polytomous test units or purely dichotomous test units.

Summarily, this study examined a 4 (proportion of polytomous test units) by 2 (total points) by 2 (MST structures) by 2 (routing module designs), resulting in a total of 32 conditions. The outcome measures for this research included several indices of measurement precision for each individual examinee for mixed-format MST. For each of the 32 conditions, 100 replications that include 1,000 simulated examinees were performed.

Test Unit Pool

A mixed-format test unit pool, the science for the 1996 National Assessment of Educational Progress (NAEP), was used for this study. A total of 424 test units contain 244 dichotomous test units, 113 test units having two-step difficulties, and 67 test units having three-step difficulties. The test unit pool also contains three content areas: physical science test (29.72%); earth science test units (34.90%); and 150 life science test units (35.38%). Test units are calibrated according to GPCM. Since the NAEP assessment is a low-stake test, the discrimination parameters for all test units are added by a constant 0.40 (Burt, Kim, Davis, & Dodd, 2003; Grady & Dodd, 2009). Each pathway in MST mirrored the three content area and their proportions in the current pool.

Test Unit Type Proportions

Unlike previous studies that varied the amount of information of the routing modules (e.g., Jodoin et al., 2006; Kim et al., 2012; Zenisky, 2004), the current study systematically varied the test unit type distribution in the routing module.

According to the design condition, each pathway in MST should provide total points of either 40 or 60, while satisfying the proportions of the test unit types. Therefore, the construction must consider the distribution of test unit types within a specific pathway. For instance, under the condition of 60 total points and 30 percent for polytomous test units, 42 points (i.e., 70 percent of the total points of 60) should come from test units having a one-step difficulty and 18 points should come from test units having two- or three-step difficulties. The distribution of points for test units having two- and three-step difficulties was performed by considering the proportion between these two test unit types within the pool. That is, more test units with two-step difficulties were included in the form since this pool provides approximately twice as many test units of two-step difficulties (i.e., 113 test units) than test units of three-step difficulties (i.e., 67 test units). Therefore, a possible point distribution for the test is 42, 12, and 6 for the one-, two- and three-step difficulties respectively. Thus far, the test unit points within a pathway are determined. To fully specify a MST, the test unit points within the modules still need to be determined.

Routing Module Design

Two methods for test unit type distribution within the routing module are presented here. The first method constructs the routing module by using the mixture of dichotomous and polytomous test units, conveniently called the mixed routing module method. For instance, 19 points for the routing module can be distributed at 10, 6, and 3 for test units with a one-, two- and three-step difficulties respectively, rather than distributed at 19 test units of a one-step difficulty. The second approach constructs the routing module by using purely dichotomous test units, conveniently named the dichotomous routing module method. For the same example mentioned above, 19 points for the routing module were distributed at 19 dichotomous test units. Figures 3.1 through 3.4 present the potential pathway structures for the mixed routing module method in terms of various proportions and total point conditions. Figures 3.5 through 3.8 show the potential pathway designs for the dichotomous routing module method for the same proportion and total point conditions.

MST Assembly

In operational testing, test designers rely on automated test assembly (ATA) techniques to build MST forms. ATA enables test constructors to build a large number of panels from the pool while satisfying various requirements from the test blueprints. For MST, there are bottom-up, top-down, and mixed methods used to build forms (Luecht & Nungester, 2000). Bottom-up begins with module construction, and the pathways are constructed by assigning modules. Top-down methods begin with pathway construction, and the modules are constructed by allocating test units to modules within the pathway. Mixed method is a combination of both bottom-up and top-down methods. For the current study, pathway level constructions were performed. For instance, the pathway of medium difficulties (i.e., medium-medium-medium pathway) could be assembled first, followed by the easy and hard difficulty pathways for 1-3-3 design. After the construction of a medium-difficulty pathway, however, the test units for the

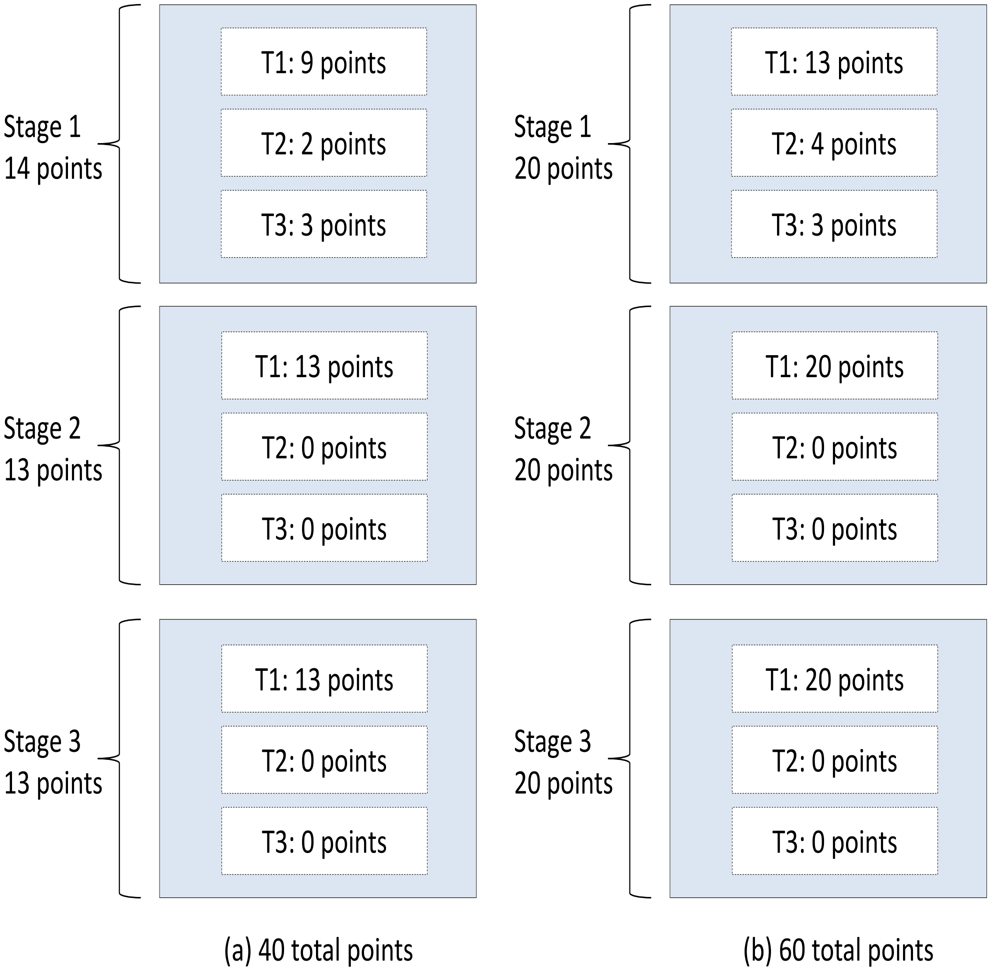

Figure 3.1: Module structures for the 10% proportion of polytomous test units and the mixed routing module method condition.

Note. T1,T2, and T3 indicate test unit types for one-, two-, and three-step difficulties.

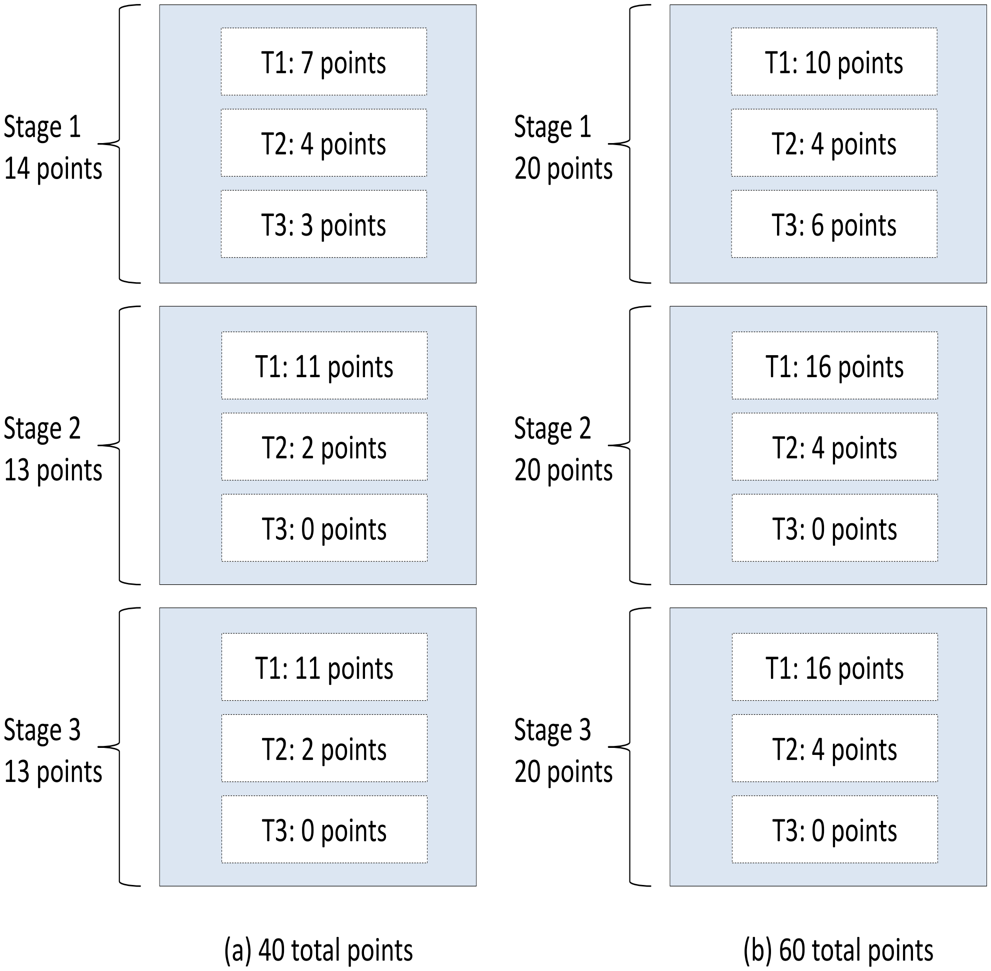

Figure 3.2: Module structures for the 30% proportion of polytomous test units and the mixed routing module method condition.

Note. T1,T2, and T3 indicate test unit types for one-, two-, and three-step difficulties.

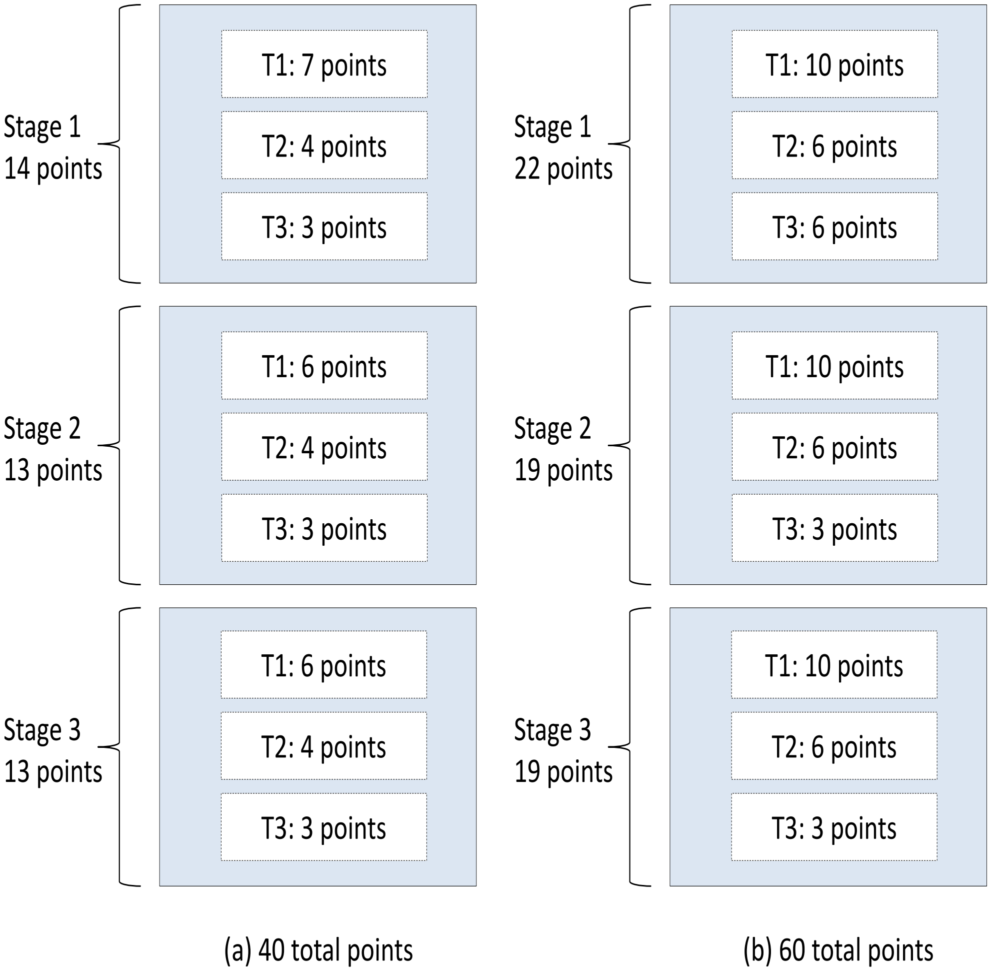

Figure 3.3: Module structures for the 50% proportion of polytomous test units and the mixed routing module method condition.

Note. T1,T2, and T3 indicate test unit types for one-, two-, and three-step difficulties.

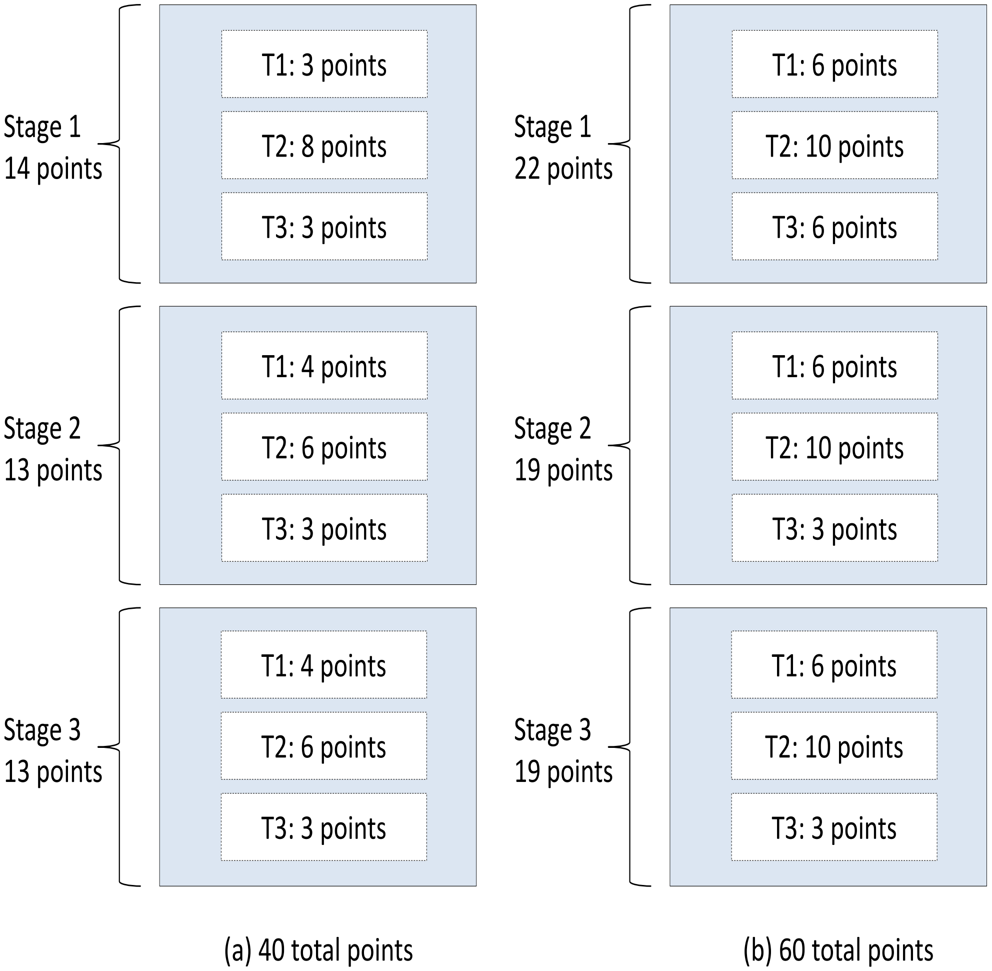

Figure 3.4: Module structures for the 70% proportion of polytomous test units and the mixed routing module method condition.

Note. T1,T2, and T3 indicate test unit types for one-, two-, and three-step difficulties.

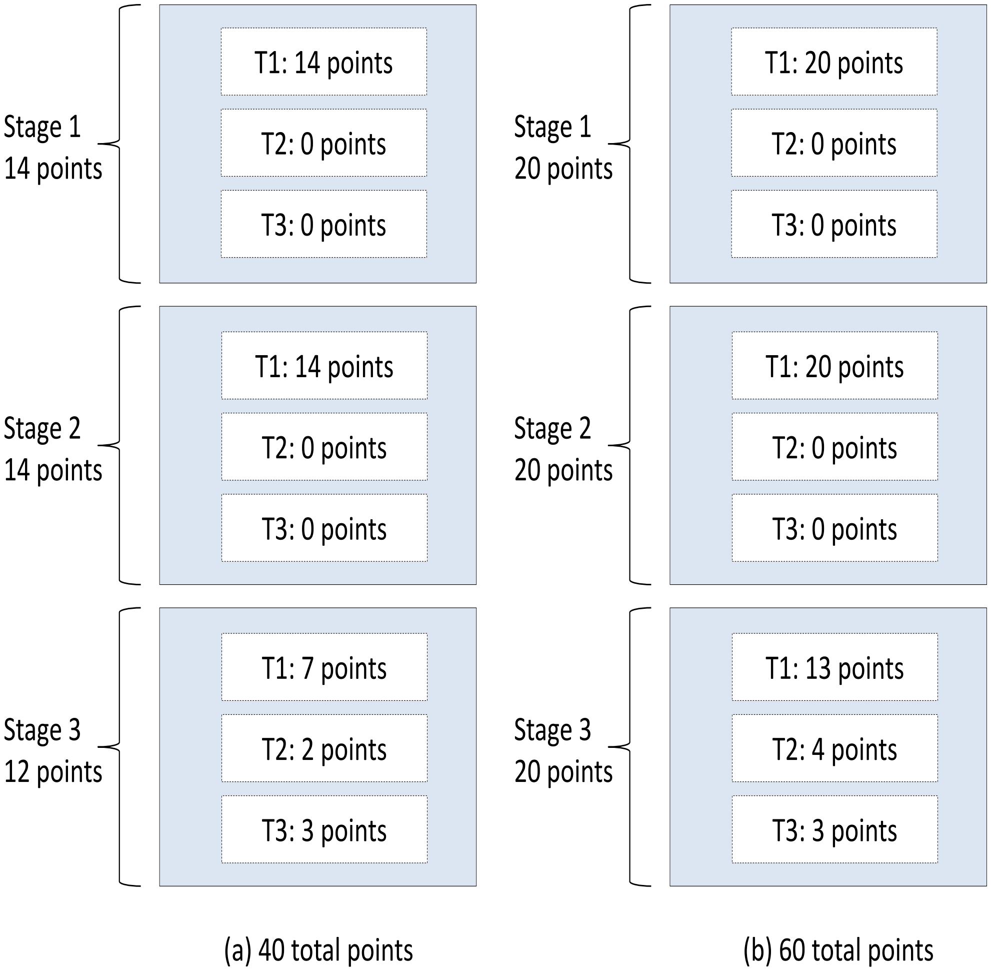

Figure 3.5: Module structures for the 10% proportion of polytomous test units and the dichotomous routing module method condition.

Note. T1,T2, and T3 indicate test unit types for one-, two-, and three-step difficulties.

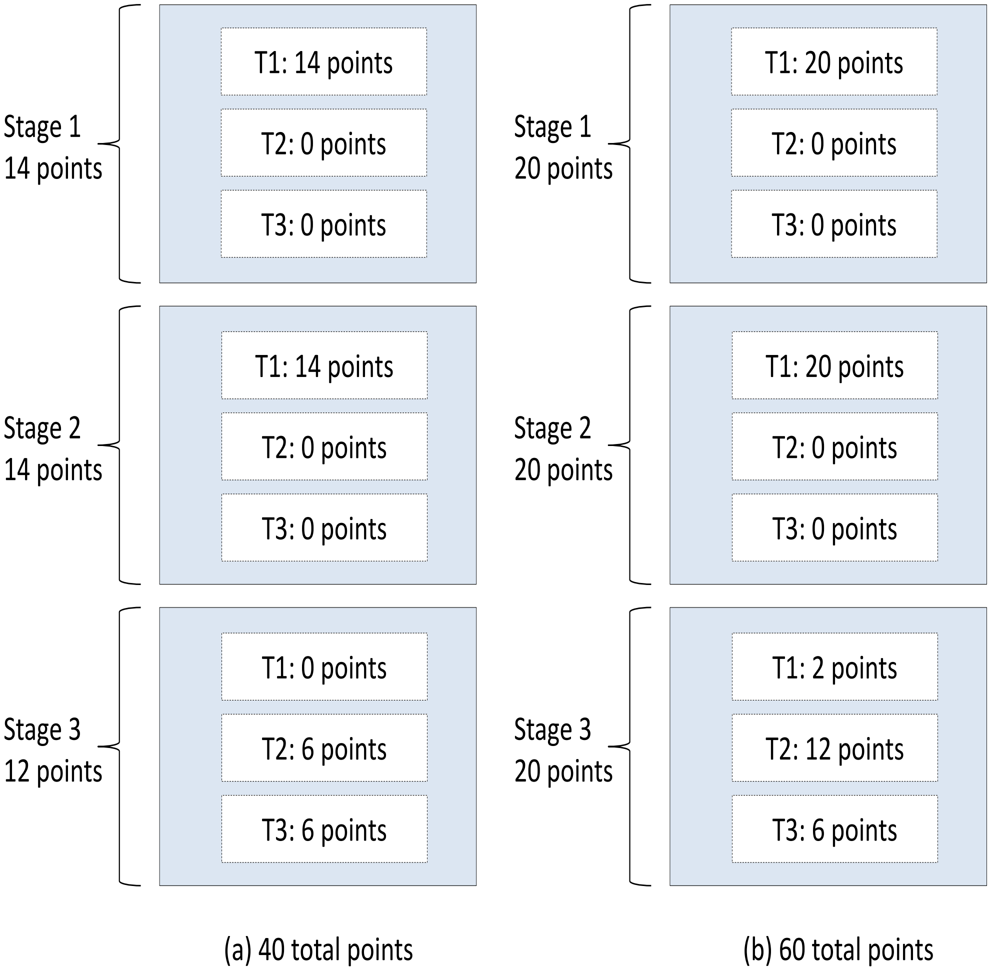

Figure 3.6: Module structures for the 30% proportion of polytomous test units and the dichotomous routing module method condition

Note. T1,T2, and T3 indicate test unit types for one-, two-, and three-step difficulties.

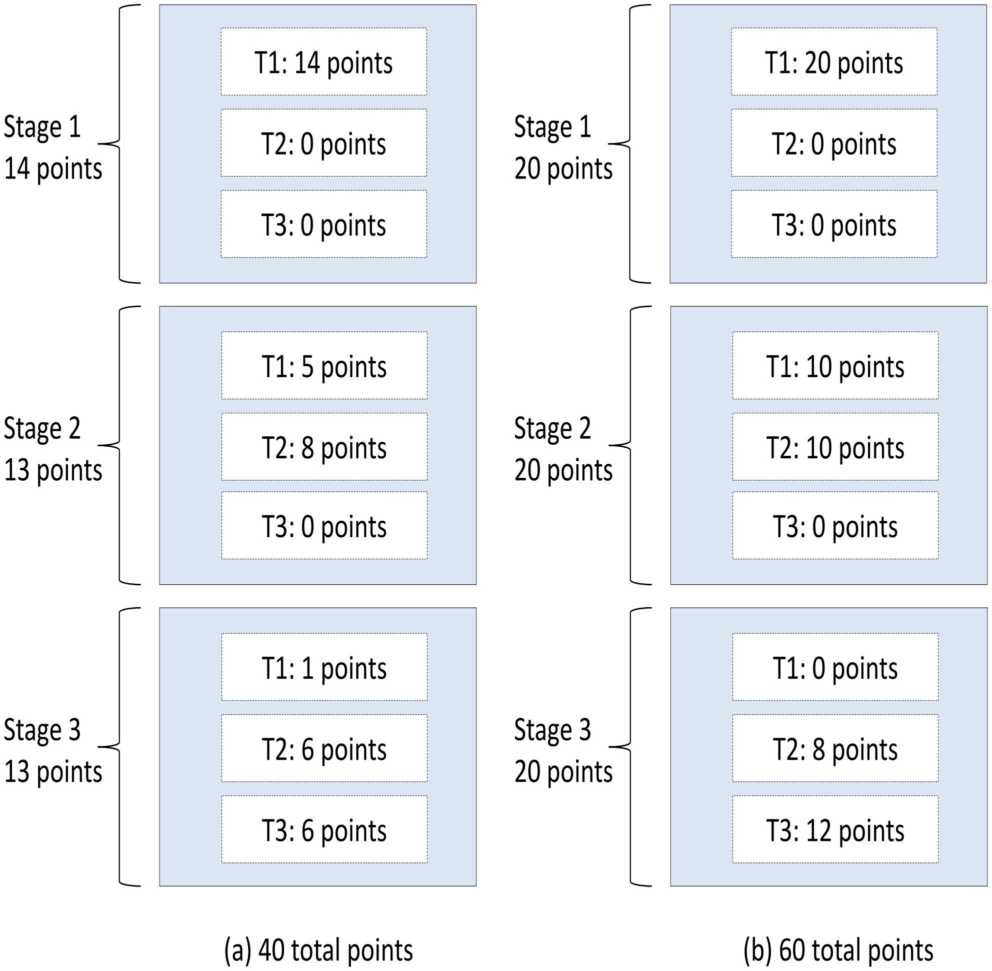

Figure 3.7: Module structures for the 50% proportion of polytomous test units and the dichotomous routing module method condition

Note. T1,T2, and T3 indicate test unit types for one-, two-, and three-step difficulties.

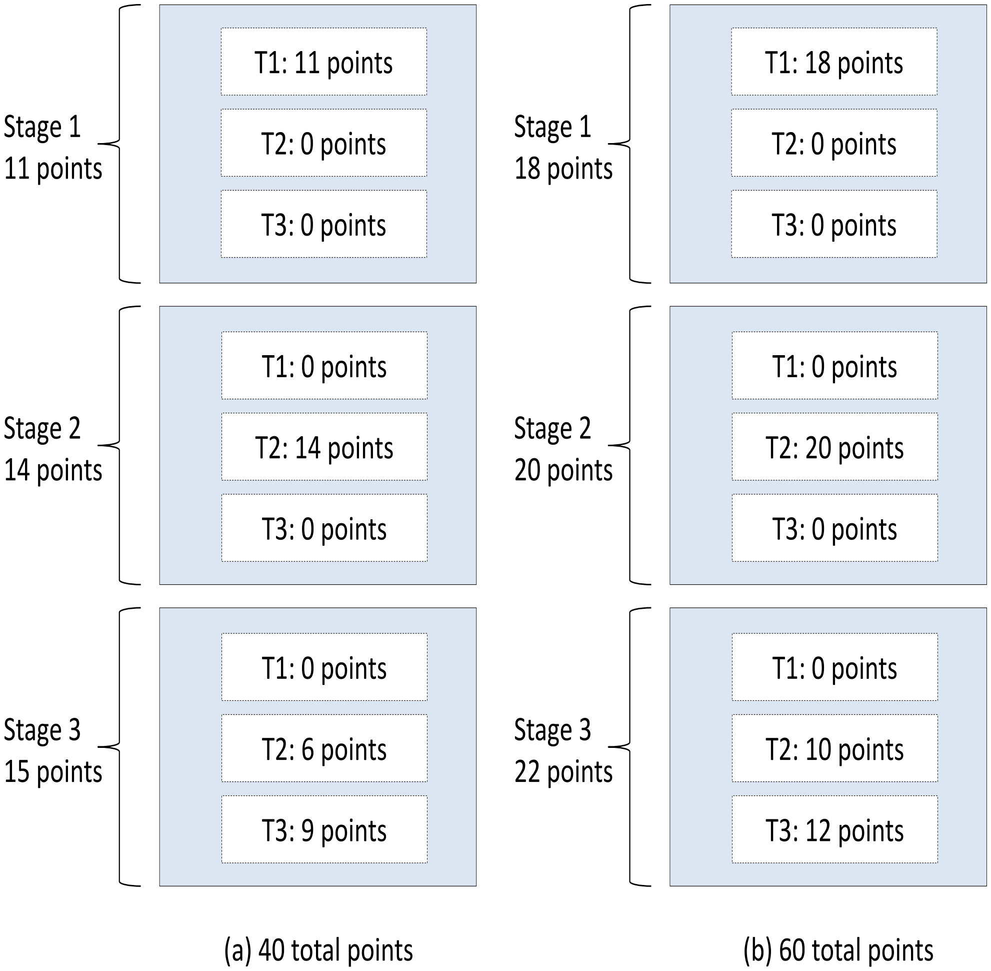

Figure 3.8: Module structures for the 70% proportion of polytomous test units and the dichotomous routing module method condition

Note. T1,T2, and T3 indicate test unit types for one-, two-, and three-step difficulties.

first module needed to be determined so that the subsequent pathways can include them during their constructions.

MST constructions were performed in R (Ihaka & Gentleman, 1996) using the lpSolveAPI package (Konis, 2011). To control the peak of test information, minimax LP models (van der Linden, 1987) were specified based on the target TIFs for each condition. The LP model specifying statistical constraints is written as:

$$minimize\, y, \tag{3.1}$$

$$\sum_{i=1}^{N} I_i(\theta_k)x_i - T(\theta_k) \leq y, for\; all\; k, \tag{3.2}$$

$$T(\theta_k) - \sum_{i=1}^{N} I_i(\theta_k)x_i \leq y, for\; all\; k, \tag{3.3}$$

$$\sum_{i=1}^{N} S(x_i) = n, \tag{3.4}$$

$$y \geq 0, \tag{3.5}$$

$$and \tag{3.6}$$

$$x_i \in \{0, 1\}, i = 1, ..., N, \tag{3.7}$$

where x_i is the decision variable for test unit i; y is the real-valued absolute deviation from target TIF; N is the number of test units; $I_i(\theta_k)$ is the information for test unit i at θ_k; $T(\theta_k)$ is the target TIF at θ_k; k is the number of theta points at the latent trait scale; $S(x_i)$ is the points of test unit i; and n is the target points in the pathway. Two panels for each condition were constructed.

Once the pathway is constructed to meet the target TIF, the test units that occupy the routing module (i.e., the module at the first stage) are then determined. Since test units within the routing module are shared among three major pathways in the 1-3-3 design, the construction of the other two pathways was performed with test units in the routing module already included. This task is easily accomplished by adding one constraint to LP model. After the construction of the first module, the test units in the routing module were included for the second and third pathway constructions utilizing the following constraint:

$$\sum_{i \in R_j} x_i = n_j, \tag{3.8}$$

where R_j is the set of routing test units for panel j that is determined from the construction of the first pathway, and n_j is the number of routing test units. Under each condition, content balancing for the three content areas was applied to a pathway level.

Target TIF

Determining reasonable targets for MST forms is one of the essential steps in MST construction. Since MST is one form of measurement tool, a test needs to be constructed to meet the purported goals of that instrument. Test assembly, as introduced by Birnbaum (1968), achieves measurement accuracy through the target TIF. For example, if a test needs to assess ability across a wide range of the theta scale, a relatively flat target TIF is preferable. On the other hand, a licensure test for making a pass/fail decision may require a target TIF that peaks around the cutoff theta score in order to increase accuracy at the cutoff point. In addition, test developers need to consider the statistical characteristics of a test unit in a given pool to determine the target TIFs that can actually be constructed (Luecht & Burgin, 2003).

In order to construct reasonable target TIFs in this study, multiple constructions were performed using 50 percent of the polytomously scored test units without the test unit replacements, while following the content balancing requirements. Target TIFs were determined for each pathway difficulty and the total points, resulting in a total of 10 targets; five pathway difficulties (i.e., 3 pathway difficulties located at -1.2, 0.0, and 1.2 on a theta scale for 1-3-3 design + 2 pathway difficulties located at -0.5 and 0.5 for the 1-2-2 design) across two total points (i.e., 40 and 60).

Data Generation

Under GPCM, the probability of responding in each category is obtained given the test unit parameters and an examinee's ability. Examinees' abilities were randomly drawn from standard normal distribution. Given the true ability, the probability of a response in each category is then computed. The cumulative probability of a response is then calculated by summing the category response probabilities. A random number drawn from a uniform distribution is then compared to the cumulative probability. The response is assigned to the smallest score for which the cumulative probability is larger than the random number. This procedure was repeated to generate 1,000 response strings for all 424 test units. In each replication, MST simulation was performed using one of the 100 data sets. Data generations were performed using R (Ihaka & Gentleman, 1996).

MST simulation

Given the response data sets and the MST designs constructed from ATA, a MST simulation was performed. One of two panels in MST form was randomly chosen and the routing module then was administered to an examinee. After the examinee finishes the routing module, the ability estimation was performed using EAP. At each stage, ability estimation was performed from all the previous test units in the MST and their respective response patterns. The relative standing of an examinee's ability estimates relative to the pathway difficulties determined the routing decision. In other words, an examinee was routed to a pathway where the distance between the current ability estimate and the pathway difficulty is the smallest. MST simulation then was performed using R (Ihaka & Gentleman, 1996).

Data Analysis

The results of MST simulation were analyzed with regard to the quality of ability estimation measured from the comparison between true θs and their estimates from MST simulation.

Accuracy of Ability Estimation

The accuracy of ability estimates is also of interest for any measurement procedure. First, descriptive statistics (e.g., mean, minimum, and maximum) of the estimated θs and the Pearson correlation between known and the estimated θs are presented for each study condition. These values were averaged across 100 replications and each replication includes 1,000 simulated examinees. Secondly, Root Mean Square Error (RMSE) and bias were calculated to evaluate the recovery of known θs. RMSE and bias are calculated as:

$$RMSE = \sqrt{\left[\frac{\sum_{j}^{N}\left(\theta_j - \hat{\theta}_j\right)^2}{N}\right]}, \tag{3.9}$$

$$Bias = \frac{\sum_{j}^{N}\left(\theta_j - \hat{\theta}_j\right)}{N}, \tag{3.10}$$

where N is the number of examinees, $\hat{\theta}_j$ is the ability estimate of examinee j, and θ_j is the true ability of examinee j. Average RMSE and bias were calculated across 100 replications for each of the study conditions.

Conditional plots (i.e., conditional on θ) for RMSE, bias and standard error (SE) averaged across 100 replications were also presented to highlight the measurement precision in various parts of the θ scale. RMSE, bias and SE were plotted on 13 data points of θ from -3.0 to 3.0 in increments of 0.5. The actual θ ranges were from -3.25 to 3.25 in order to achieve midpoints that range from -3.0 to 3.0 (e.g., a midpoint of -3.0 represents θ between -3.25 to -2.75).

For each replication, average RMSE, bias and SE were calculated for each bin (i.e., $RMSE_b$, $bias_b$ and SE_b, where b=1,2,...,13). $RMSE_b$, $bias_b$ and SE_b were averaged across 100 replications to produce the grand means of RMSE, bias and SE for the conditional plots.

MST Test Information

Equation 2.7 provides a useful tool for calculating standard errors of measurement once the test information function is known. For adaptive testing, however, deriving the test information function has not been attempted by previous researchers. One of the main reasons for this is that the number of test forms is exceedingly large due to their adaptive nature. Therefore, simulation studies have been conducted to assess various aspects of performances, such as standard errors, bias, and item utilization. Unlike CAT, MST has unique characteristics. Test forms are constructed before the administration using test construction techniques, while the adaptive points during the administration occur at the end of stage. Thus, MST could be considered to have characteristics of both the linear test and fully adaptive CAT. Unlike previous researchers who have relied exclusively on simulation results for the performance of test forms, the current study attempts to derive standard error analytically by forming test information (conveniently called MST test information) while considering aspects of both the linear and adaptive nature of MST.

The benefits of comparison between empirical and analytic standard errors are twofold. First, analytic standard errors might provide additional reference points for the simulation results. MST simulation results rely on largely three factors: the statistical quality of the pool, MST constructions, and MST administration methods. Simulation can produce different results by changing the factors necessary for simulation, such as target test information, population sampling, and replication number. These factors can interact with the study condition (e.g., proportion in the current study), meaning the interpretation of the outcome might not be straightforward. Researchers can acquire extra confidence in their results by matching simulation outcomes to analytically driven ones. Second, analytic standard errors can provide a theoretical upper limit on the MST performance. As analytic standard errors are derived directly from the test information, they represent the highest measurement accuracy one can achieve from the MST design. Simulation results that fall below this level of performance should raise questions regarding the design factors within the MST administration.

Two observations that lead to the MST test information were introduced. First, current MST practices were used to construct MST modules targeting specific difficulties. For example, modules for the second and third stages of 133 MST were constructed to cover easy, medium, and hard difficulties. As a result, major pathways were formed to represent specific difficulties, and each major pathway could be considered a linear test form sharing a common routing module. Second, following previous research (Kim et al., 2013; Zenisky, 2004), a large portion of examinees were routed through major pathways. Zenisky considered MST performance in the context of certification and licensure assessment using four routing methods: DPI, proximity, routing based on number-correct scores (NC), and random routing. The study results showed that the average proportion of examinees routed through major pathways for the three routing methods (i.e., DPI, proximity, and NC) were 81.74%, 81.6%, and 71.15% for the 133 MST design and 90.13%, 79.8%, and 81.34% for the 122 MST design. Kim et al. compared panel designs with a routing method using the partial credit model and reported that 73% and 78% of examinees were routed through major pathways for the AMI and DPI routing methods, respectively. Thus, once examinees abilities are estimated at the routing module, a predominant portion of them do not change the difficulty level of the module in the subsequent stages. In other words, the most important adaptivity in MST design occurs at the routing module.

MST test information was derived from two observations: the linearity of major pathways and the adaptivity at the routing module. As adaptivity occurs depending on the ability estimates, MST test information is formed by adaptively choosing major pathway information functions. A logical method was to select the range of pathways to which examinees of abilities are most likely routed.

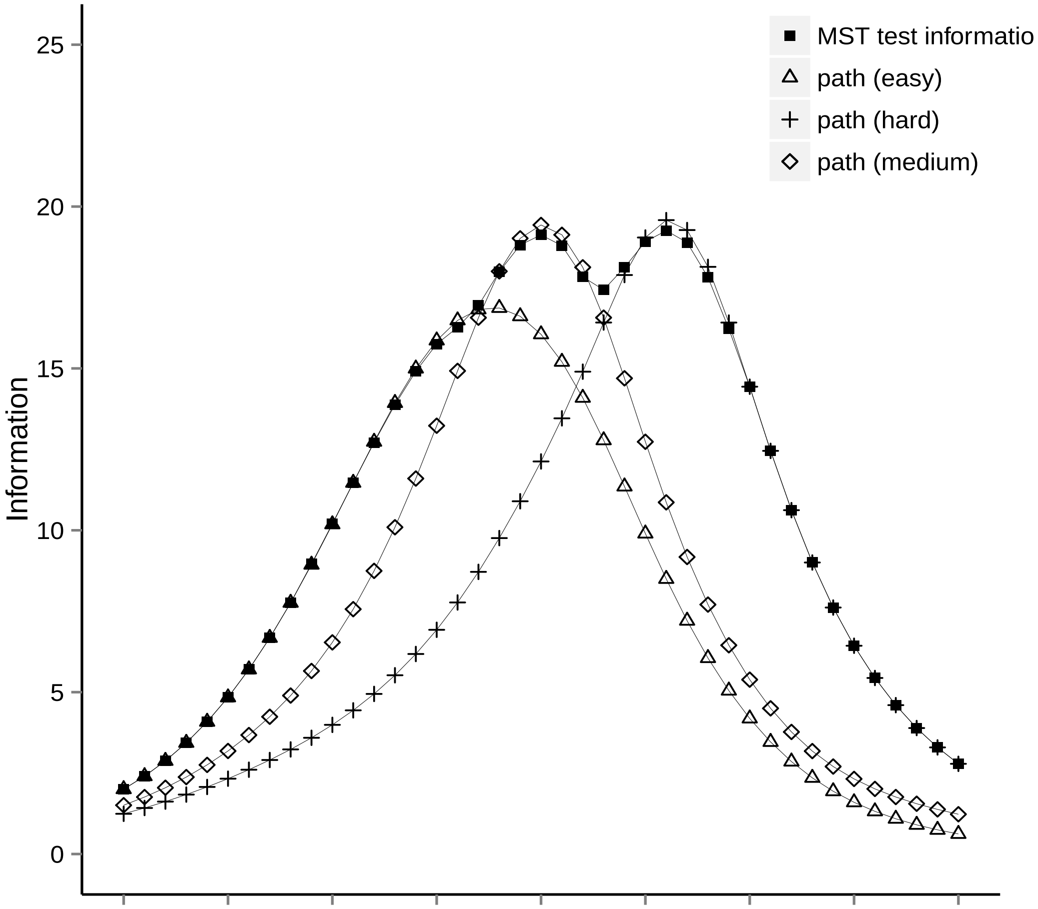

Figure 3.9: MST test information derived from major path information functions.

Formally, the MST test information for 133 MST is expressed as:

Figure 3.9 shows an example of the 133 MST test information function constructed from three pathway information functions. Easy, medium, and hard difficulty pathways were constructed so that their difficulties were located at -1.2, 0.0, and 1.2 on a theta scale, while the easy and medium difficulty pathways intersected at theta = -0.6 and the medium and hard difficulty pathways intersected at theta = 0.6. Therefore, the MST test information function was constructed from the easy pathway, where theta is less than -0.6; the medium pathway, where theta is between -0.6 and 0.6; and the hard pathway, where theta is larger than 0.6.

Formally, the MST test information for 133 MST is expressed as:

$$\begin{aligned} TI_{mst}(\theta) = & 1(\theta < C_1)PI_{easy} + \\ & 1(C_1 <= \theta <= C_2)PI_{medium} + \\ & 1(\theta > C_2)PI_{hard} \end{aligned} \tag{3.11}$$

, while the MST test information for 122 MST is expressed as:

$$TI_{mst}(\theta) = 1(\theta < C_1)PI_{easy} + 1(\theta >= C_1)PI_{hard} \tag{3.12}$$

where 1(*condition*) is 1 if the *condition* is true and zero otherwise; PI is the pathway information function for each difficulty; and C_1 and C_2 are midpoints of two adjacent major pathway difficulties.

Equations 3.11 and 3.12 were used to calculate the MST test information for a panel in MST. When multiple panels were constructed, the overall MST test information was the average of the MST test information functions calculated using Equations 3.11 or 3.12 for each panel. The MST test information functions were then smoothed using a moving average to achieve realistic test information and the standard error functions. The moving average of 5 points for $TI_{mst}(\theta)$, for example, is expressed as:

$$TI_{mst}(\theta) = \frac{\sum_{i=1}^{5} TI_{mst}(\theta + i - 3))}{5} \tag{3.13}$$

The standard error of measurement was then calculated using Equation 3.14, which is the inversed squared root of the MST test information function.

$$SE_{mst}(\theta) = \frac{1}{\sqrt{TI_{mst}(\theta)}} \tag{3.14}$$

Chapter 4

RESULTS

Overview

This chapter consists of two sections detailing (1) MST constructions and (2) MST simulation results. The first section presents the outcome of the MST test constructions by focusing on the MST test information functions for 32 study conditions - namely, four proportions of the polytomous test units, two total points, two MST panel structures, and two routing module designs. The second section provides details on the MST simulation results. For each condition, the mean Pearson correlation between the known and estimated thetas, the descriptive statistics of the estimated thetas, and the frequency of the use of each pathway are presented. The averages of RMSE and bias are also presented, and the conditional grand mean standard errors and mean bias plots for each condition are plotted.

MST Constructions

This section presents the test unit pool information and the actual construction of MST through the MST test information functions. Test unit pool information provides overall distribution of information that the pool can provide. The analysis of the test unit pool information is crucial in setting a realistic goal of measurement precision that a test can achieve across ability levels.

Test unit pool information

The test unit pool information function (i.e., the accumulation of test unit information in the pool) is presented in Figure 4.1. The pool information was slightly negatively skewed, and there was relatively little information at the easy difficulty levels. In addition, the location of the peak is at the boundary of hard and medium difficulty levels, indicating that a large amount of information is focused at thetas between zero and 2.0.

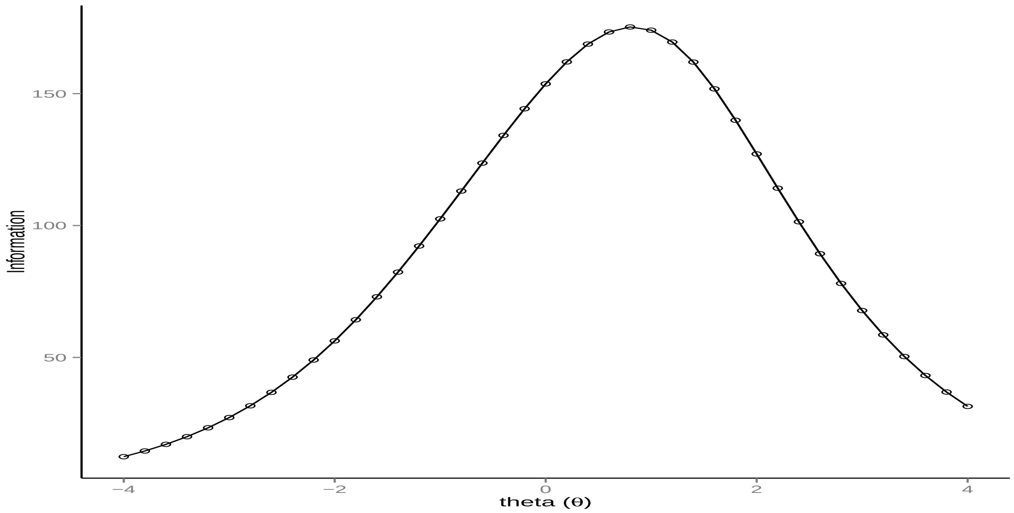

Figure 4.1: Information function of test unit pool.

MST test information function

MST test information was derived from concatenating major pathway information to approximate the test information of MST forms. As all MST forms contain two panels for a given condition, the two MST test information functions were averaged. For ease of illustration, the theta scale was divided into easy, medium, and hard levels. The boundary between easy and medium was located around theta = -1.0, and the medium and hard levels were separated by theta = 1.0.

Figures 4.2 and 4.3 present the MST test information for the 10% proportion of polytomous test units for both the dichotomous and mixed routing module conditions. The peak information values of MST test information for 60 and 40 total point conditions were 20 and 15 respectively, which correspond to the values specified in target test information functions. In terms of MST structure, MST with the 133 structure maintained high information for a wide range of difficulty levels compared to MST with the 122 structure. The difference between the 133 and 122 structures was large in the hard difficulty levels, while the difference in information functions was minimal between the two MST structures for easy and medium difficulty levels.

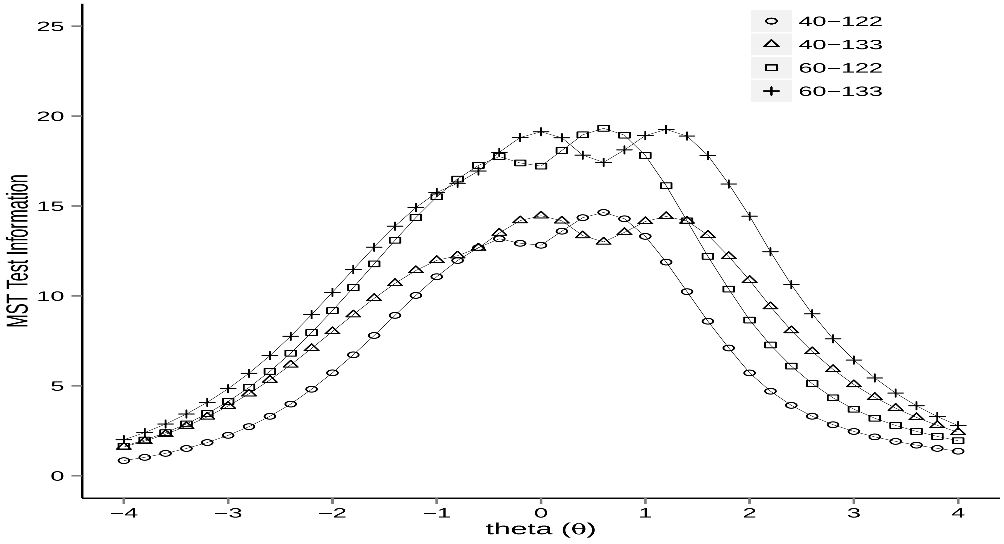

Figure 4.2: MST test information for 10% proportion of polytomous test units and dichotomous routing module condition.

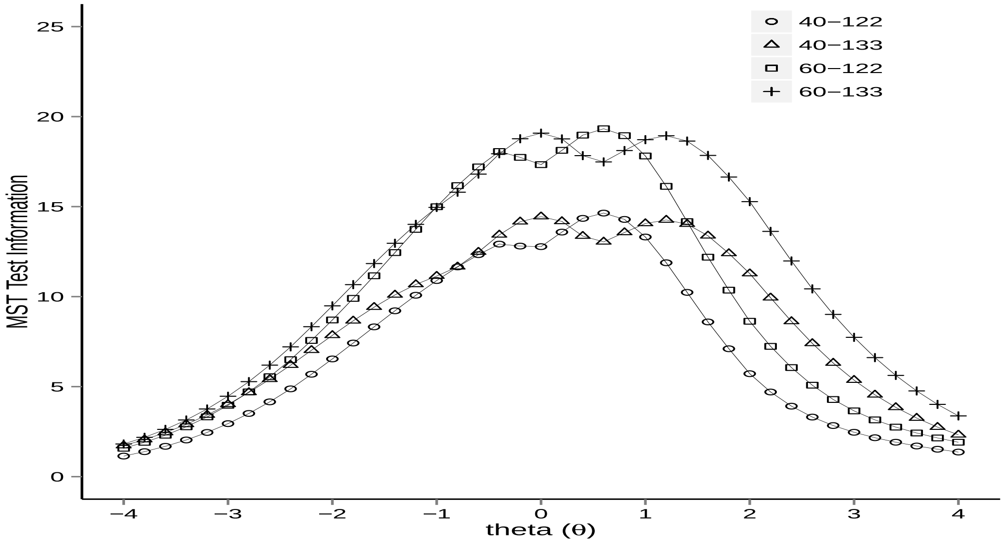

Figure 4.3: MST test information for 10% proportion of polytomous test units and mixed routing module condition.

Figures 4.4 and 4.5 present the MST test information for the 30% proportion of polytomous test units for the dichotomous and mixed routing module conditions, respectively. MST constructions performed well according to the targets in terms of the peak values and locations. The condition of 60 total points resulted in overall larger information than the 40 total points condition. In addition, the 133 structure possessed larger information than 122 in the hard difficulty levels for both total point conditions.

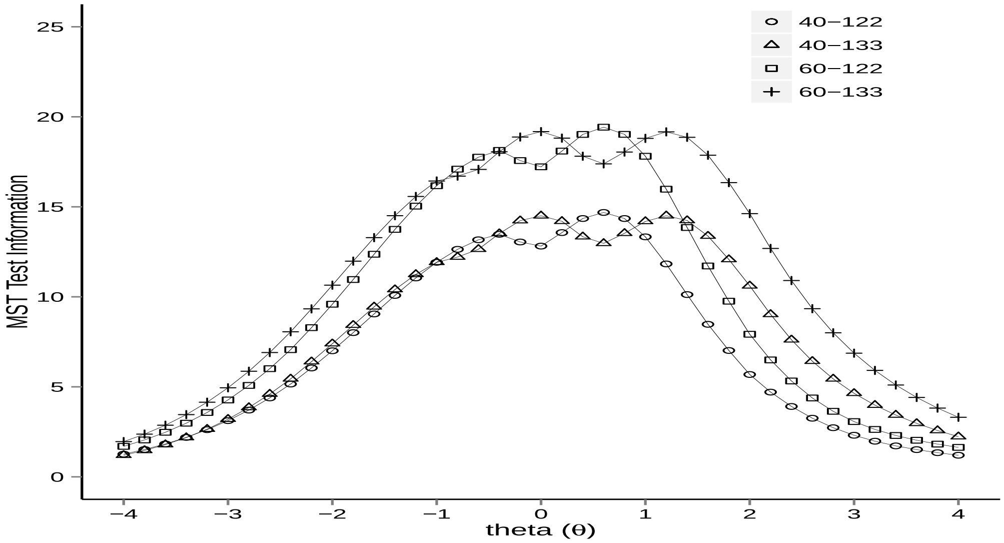

Figure 4.4: MST test information for 30% proportion of polytomous test units and dichotomous routing module condition.

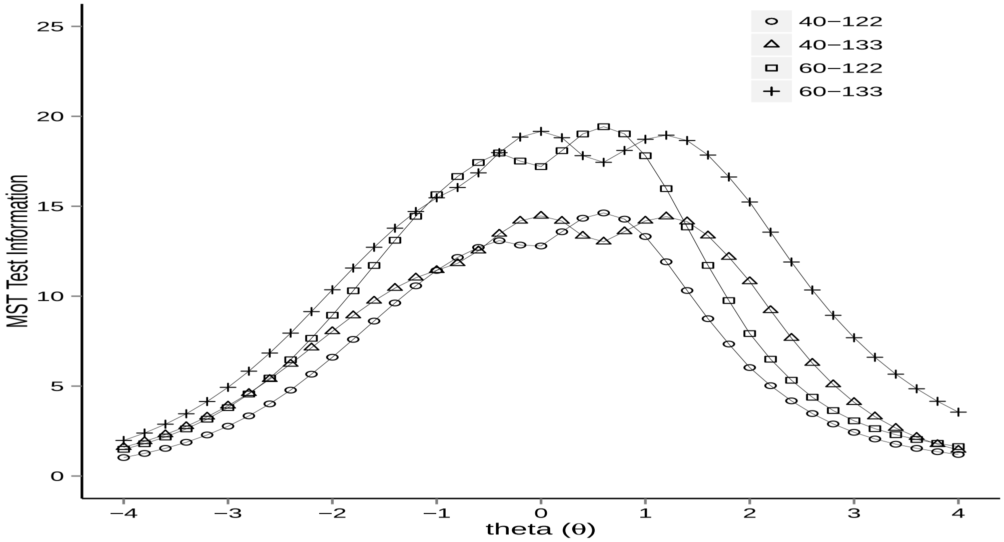

Figure 4.5: MST test information for 30% proportion of polytomous test units and mixed routing module condition.

Figures 4.6 and 4.7 present the MST test information for the 50% proportion of polytomous test units for the dichotomous and mixed routing module conditions, respectively. Similar bell-shaped information functions were observed, while peak values, peak locations, and the pattern of spread were constructed according to the specifications in targets. The information difference between the 122 and 133 structures were more pronounced in hard difficulty levels than easy difficulty levels.

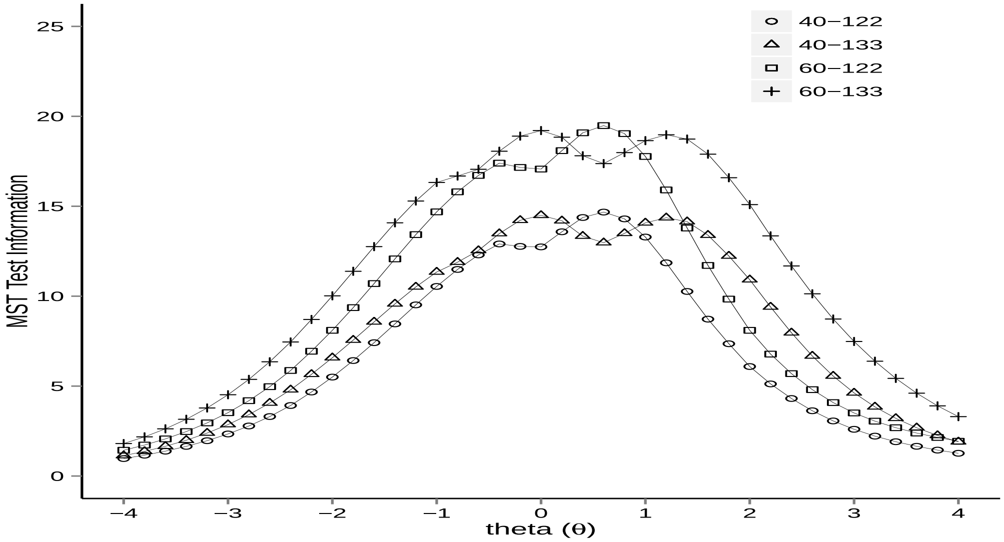

Figure 4.6: MST test information for 50% proportion of polytomous test units and dichotomous routing module condition.

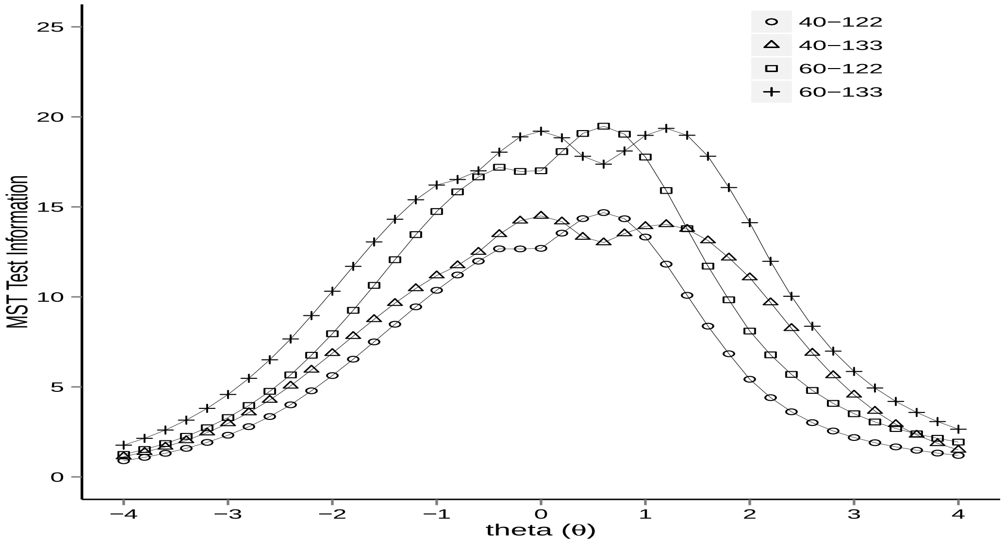

Figure 4.7: MST test information for 50% proportion of polytomous test units and mixed routing module condition.

Figures 4.8 and 4.9 present the MST test information for the 70% proportion of polytomous test units for the dichotomous and mixed routing module conditions, respectively. Compared to the other proportion conditions, the amount of information for the easy and hard levels was somewhat low, resulting in narrow MST test information functions. The information functions in the easy difficulty levels were relatively low compared to those in the hard difficulty levels. However, the MST with a mixed routing module condition provided slightly more information for easy difficulty levels than the MST with the dichotomous routing module design.

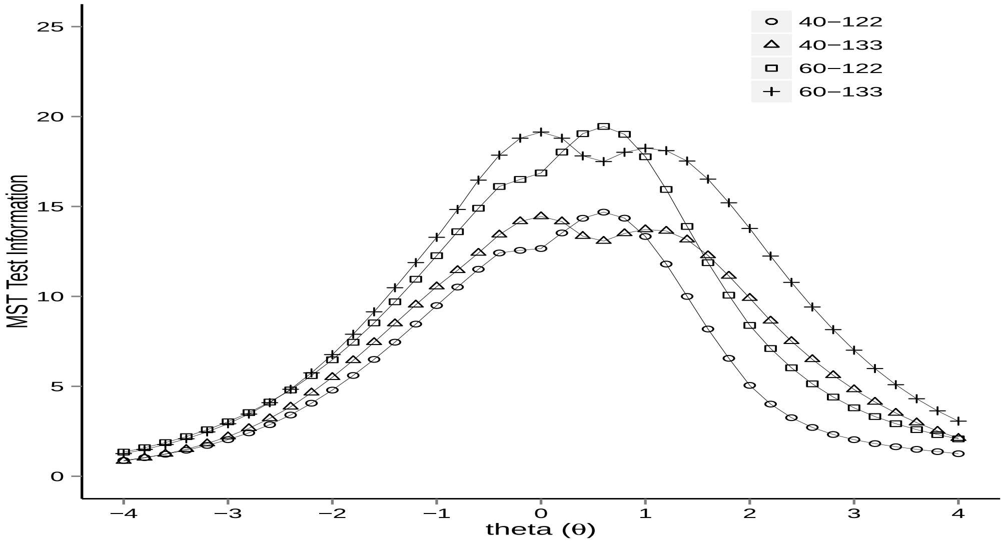

Figure 4.8: MST test information for 70% proportion of polytomous test units and dichotomous routing module condition.

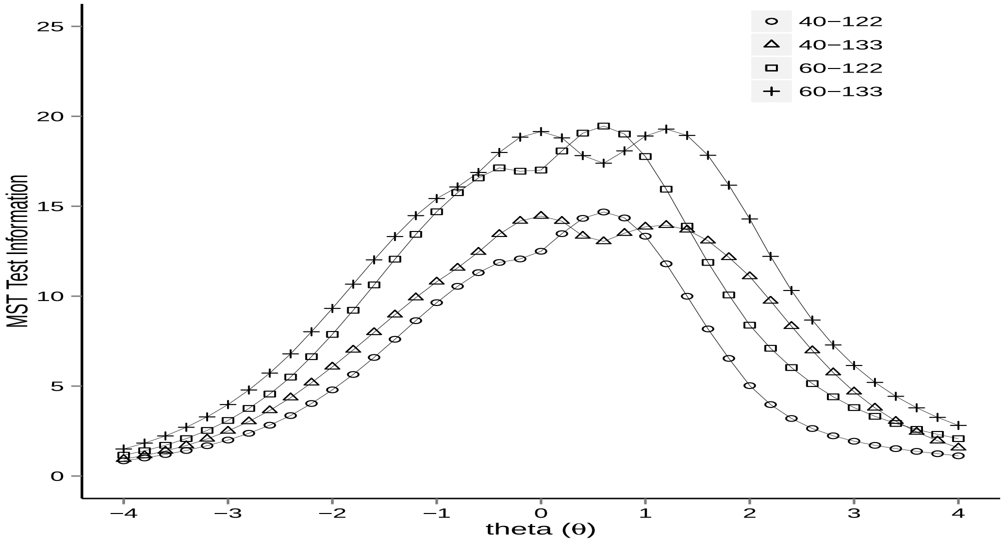

Figure 4.9: MST test information for 70% proportion of polytomous test units and mixed routing module condition.

In summary, MST constructions for all conditions followed the targets in general but reflected the characteristics of the test unit pool. The MST test information functions resulted in a bell-shaped curve showing large information in the medium range and low information as theta moves toward to either extreme. MST with 60 total points resulted in larger information than the 40 total points condition, as specified in the target test information. In terms of the MST structure, MST test information functions from the 133 condition provided larger information for a wide range of difficulty levels compared to MST with the 122 structure condition.

The routing module design and the proportion of polytomous test units also affected test construction. For example, in the 70% of proportion of polytomous test units and dichotomous routing module design, 44 of the 60 total points and 29 of the 40 total points within a pathway had to be drawn from polytomous test units (see Figure 3.8). As the routing module was composed of purely dichotomous test units, the test points drawn from polytomously scored test units were multiplied by the number of pathways and the number of panels in a MST form. For example, in the 133 MST form, polytomous items should account for 264 (i.e., 44(points) by 3(pathways) by 2(panels)) and 174 (i.e., 29(points) by 3(pathways) by 2(panels)) points for 60 and 40 total point conditions, respectively. This is the most stringent design condition requiring a large number of polytomous test units within the MST design. The actual construction demonstrated that the information for both easy and hard pathways did not meet the specified target information.

In addition, the characteristics of the test unit pool were reflected in the MST test information. For easy difficulty pathways, the amount of information at the peak was lower than the target (i.e., between 75 and 80% of the target value), forming unsymmetrical MST test information functions for all 32 study conditions. This relatively low information in the easy level was more pronounced for the 133 design because the easy pathways in the 133 MST structure was located at further negative theta points (i.e., theta = -1.2) than in the 122 MST designs (i.e., theta = -0.6).

Figure 4.10 shows the actual construction of MST test information of MST with the 70% of proportion of polytomous test units, 133 structure and dichotomous routing module design compared to the target test information functions. For each pathways, the actual construction could not provide enough information to match the target information functions. The peak values were around 11.6 (the target information value of 20) and 10.9 (the target information value of 15) for MST with the 60 and 40 total point conditions respectively, preventing the formation of unique peaks.

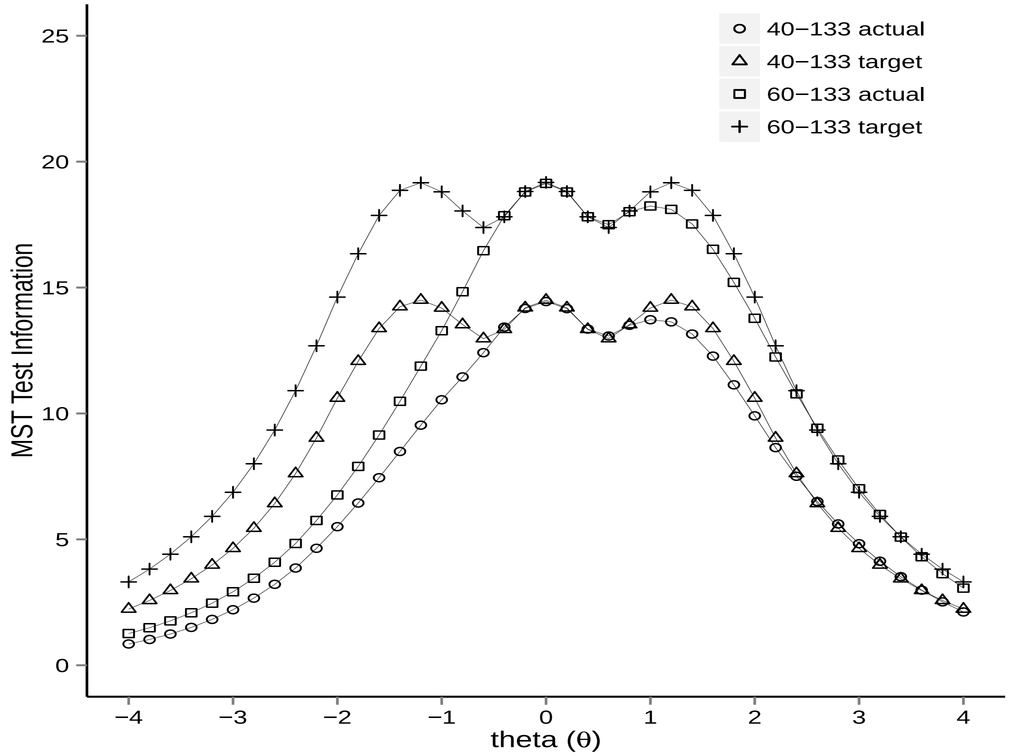

Figure 4.10: Target test information and actual MST test information functions for 70% proportion of polytomous test units, 133 structure and dichotomous routing module condition.

MST Simulation Results

Descriptive statistics

Table 4.1 provides the averaged descriptive statistics of estimated θs and the mean Pearson correlation between known and estimated θ averaged across 100 replications. The descriptive statistics include the grand means of θ estimates, which are the mean estimated θs averaged across the 100 replications. The grand means of θ estimates were ranged from -0.051 and -0.004 and the mean Pearson correlations between the known and estimated thetas for all conditions were equal to or above 0.96. There were not noticeable differences in mean Pearson correlation values across study conditions.

Table 4.1:
Descriptive Statistics of the Estimated θs Averaged across 100 Replications

Points	Structure	Proportion	Routing	Grand Mean	(min	max)	Mean Correlation
40	122	10	dich	0.002	(-0.095	0.091)	0.963
			mixed	-0.005	(-0.119	0.083)	0.971
		30	dich	0.002	(-0.102	0.090)	0.965
			mixed	-0.007	(-0.118	0.071)	0.970
		50	dich	-0.001	(-0.093	0.096)	0.974
			mixed	-0.006	(-0.108	0.083)	0.977
		70	dich	-0.001	(-0.097	0.086)	0.974
			mixed	-0.009	(-0.106	0.087)	0.977
	133	10	dich	0.004	(-0.096	0.093)	0.965
			mixed	-0.005	(-0.109	0.089)	0.969
		30	dich	0.003	(-0.101	0.092)	0.965
			mixed	-0.009	(-0.117	0.071)	0.969
		50	dich	0.000	(-0.100	0.084)	0.973
			mixed	-0.005	(-0.114	0.086)	0.978
		70	dich	-0.001	(-0.094	0.091)	0.973
			mixed	-0.008	(-0.117	0.077)	0.978
60	122	10	dich	-0.008	(-0.106	0.088)	0.963
			mixed	-0.012	(-0.117	0.077)	0.968
		30	dich	0.000	(-0.096	0.105)	0.963
			mixed	-0.015	(-0.115	0.076)	0.968
		50	dich	-0.006	(-0.117	0.087)	0.973
			mixed	-0.007	(-0.123	0.074)	0.978
		70	dich	-0.002	(-0.118	0.092)	0.973
			mixed	-0.006	(-0.114	0.084)	0.977
	133	10	dich	-0.038	(-0.123	0.056)	0.960
			mixed	-0.031	(-0.136	0.055)	0.966
		30	dich	-0.051	(-0.150	0.047)	0.960
			mixed	-0.036	(-0.142	0.058)	0.967
		50	dich	-0.012	(-0.120	0.084)	0.971
			mixed	-0.018	(-0.124	0.074)	0.974
		70	dich	-0.008	(-0.118	0.101)	0.973
			mixed	-0.010	(-0.116	0.083)	0.977

Table 4.2 shows the mean RMSE and mean bias (along with their minimum and maximum values) for all 32 study conditions. The mean RMSEs ranged from 0.330 to 0.473. There were a noticeable impact of total points and MST structure conditions to the mean RMSEs; the mean RMSE was smaller for MST with the 60 total points over MST with the 40 total points and the MST with the 133 structure over the MST with the 122 structure. The results showed that the mean RMSEs for 40 and 60 total point conditions were 0.427 and 0.369, respectively, while those for the 122 and 133 MST structures were 0.429 and 0.368, respectively. The mean RMSEs for the proportion of polytomous test units and routing design conditions were identical to two decimal places.

For all 32 conditions, mean biases were similarly small. The range of biases was also similar across conditions while minimum and maximum mean biases were -0.144 and 0.130, respectively.

Table 4.2: Mean RMSE and Mean Bias of the Estimated θ Averaged across 1,000 Replications

Points	Structure	Proportion	Routing	Mean RMSE (min,max)		Mean Bias (min,max)	
40	122	10	dich	0.473	(0.270, 0.829)	-0.001	(-0.093, 0.099)
			mixed	0.453	(0.267, 0.827)	-0.005	(-0.107, 0.099)
		30	dich	0.451	(0.270, 0.778)	-0.009	(-0.124, 0.096)
			mixed	0.440	(0.270, 0.741)	-0.004	(-0.106, 0.101)
		50	dich	0.453	(0.270, 0.714)	0.006	(-0.071, 0.115)
			mixed	0.468	(0.270, 0.796)	-0.007	(-0.144, 0.092)
		70	dich	0.472	(0.269, 0.708)	0.013	(-0.073, 0.130)
			mixed	0.471	(0.272, 0.774)	0.020	(-0.051, 0.114)
	133	10	dich	0.381	(0.265, 0.632)	0.004	(-0.057, 0.075)
			mixed	0.387	(0.268, 0.613)	0.006	(-0.060, 0.076)
		30	dich	0.387	(0.269, 0.608)	0.007	(-0.032, 0.073)
			mixed	0.406	(0.269, 0.663)	0.002	(-0.081, 0.081)
		50	dich	0.401	(0.266, 0.619)	0.010	(-0.055, 0.106)
			mixed	0.407	(0.266, 0.630)	0.010	(-0.055, 0.096)
		70	dich	0.391	(0.268, 0.593)	0.019	(-0.023, 0.103)
			mixed	0.391	(0.264, 0.614)	0.024	(-0.016, 0.121)
60	122	10	dich	0.388	(0.235, 0.682)	-0.005	(-0.087, 0.078)
			mixed	0.389	(0.232, 0.698)	-0.006	(-0.100, 0.077)
		30	dich	0.413	(0.234, 0.802)	-0.011	(-0.122, 0.057)
			mixed	0.415	(0.235, 0.791)	-0.011	(-0.131, 0.061)
		50	dich	0.384	(0.232, 0.559)	0.002	(-0.072, 0.099)
			mixed	0.388	(0.234, 0.652)	0.000	(-0.064, 0.076)
		70	dich	0.405	(0.232, 0.617)	0.006	(-0.079, 0.099)
			mixed	0.389	(0.233, 0.627)	0.003	(-0.048, 0.074)
	133	10	dich	0.340	(0.232, 0.564)	0.003	(-0.051, 0.062)
			mixed	0.341	(0.232, 0.539)	0.006	(-0.039, 0.065)
		30	dich	0.330	(0.232, 0.514)	0.003	(-0.037, 0.055)
			mixed	0.329	(0.235, 0.497)	0.006	(-0.032, 0.057)
		50	dich	0.335	(0.232, 0.500)	0.003	(-0.034, 0.052)
			mixed	0.339	(0.234, 0.537)	0.002	(-0.053, 0.060)
		70	dich	0.365	(0.232, 0.591)	0.014	(-0.031, 0.089)
			mixed	0.345	(0.234, 0.552)	0.006	(-0.047, 0.072)

Note. Proportion of polytomous test units.

Table 4.3 is the summary of the average pathway frequencies for structure and total points conditions. The average frequency of major pathways, where examinees do not change difficulties for the second and third stages, was as large as 92.93% for 60 total points and 87.66% for the 40 total points condition. This indicates that once the difficulty is determined at the end of the routing module, it does not change at the third stage in the MST.

Table 4.3: Average Frequency of Pathways

Structure	total points	mee	mem	mme	mmm	mmh	mhm	mhh	major pathways
122	40	46.23%	3.91%	3.66%	46.20%	na	na	na	92.43%
	60	46.49%	3.66%	3.41%	46.44%	na	na	na	92.93%
133	40	28.61%	3.46%	2.99%	29.99%	2.92%	3.48%	28.55%	87.15%
	60	28.70%	3.35%	2.81%	30.45%	2.86%	3.31%	28.51%	87.66%

.

Note. m, e and h indicate medium, easy and hard difficulty modules respectively.

Conditional bias

Figures 4.11, 4.12, 4.13 and 4.14 show the mean conditional bias for four factors in the study (i.e., proportion of polytomous test units, total points, MST structures, and routing designs) on θ. For all conditions, conditional mean bias plots are a monotonic decreasing function crossing zero in the center of the theta continuum. Small bias values at the center of the ability scale imply that known thetas were recovered accurately. On the other hand, bias values at both extremes indicate that thetas were slightly overestimated at the high ability levels and underestimated at the low ability levels.

Figure 4.11 shows the mean conditional biases for the four levels of the proportion of polytomous test units in the study. The four conditional bias plots follow the overall trend of monotonic decreasing functions, and the differences among them are small.

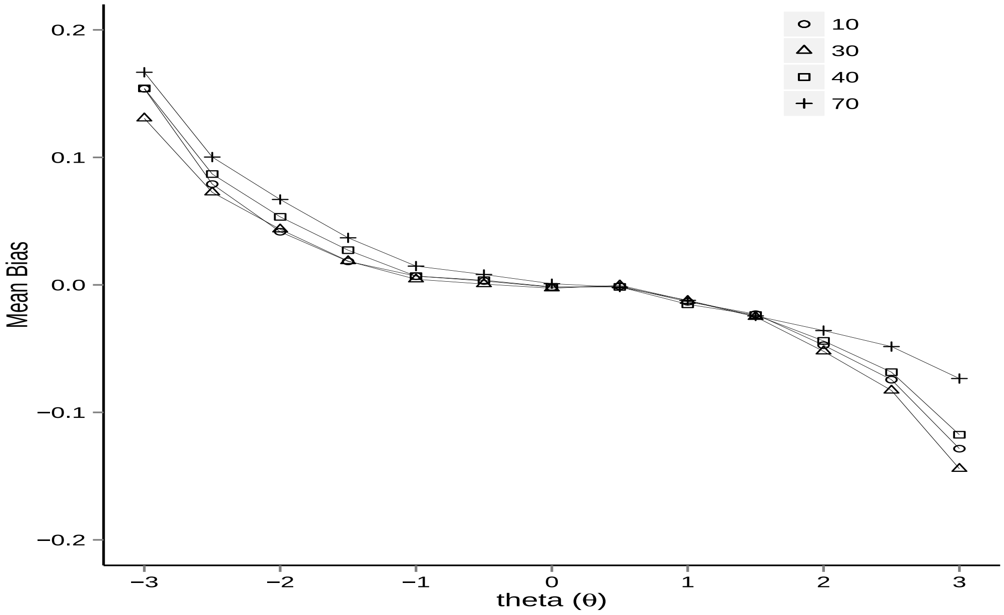

Figure 4.11: Conditional mean biases for four levels of proportion of polytomous test units

Figure 4.12 shows the mean conditional biases from two routing module variations. The two conditions produced almost identical results, and the curves appear to be on top of each other. Figure 4.13 compares the mean biases for two MST structures. The results show that the MST with the 122 structure had larger biases than the MST with the 133 structure in both extremes (i.e., overestimation at the high ability levels and underestimation at the low ability levels). Finally, Figure 4.14 shows that the MST with the 40 total points resulted in slightly larger absolute biases compared to the MST with the 60 total points condition.

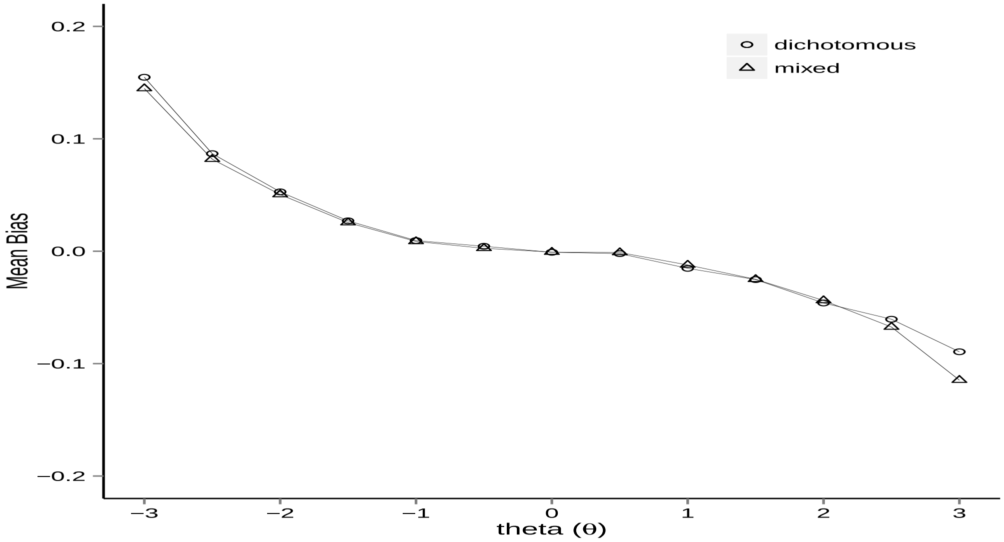

Figure 4.12: Conditional mean biases for two levels of routing module design

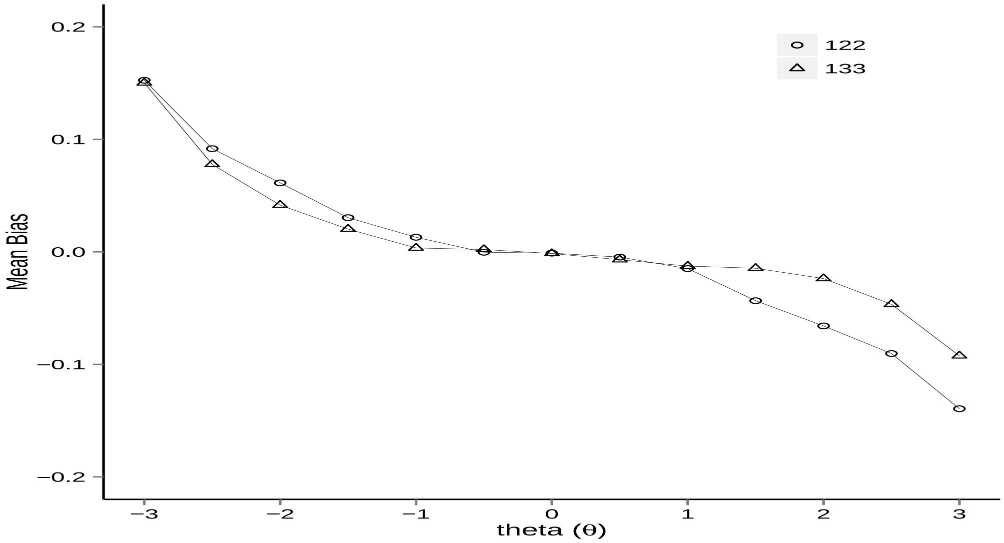

Figure 4.13: Conditional mean biases for 133 and 122 MST structure

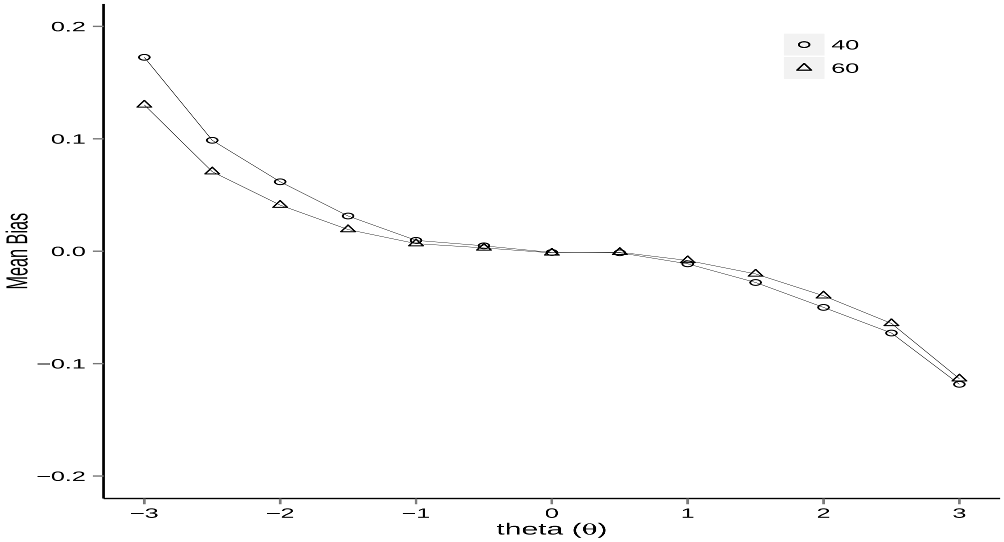

Figure 4.14: Conditional mean biases for two levels of total points

Grand mean conditional standard errors

Figure 4.15 provides the grand mean conditional standard error, which is averaged across all 32 design conditions. A typical U-shaped curve appeared, in which measurement precision is greater in the medium difficulty levels. The minimum average standard error was 0.251 around theta value of 0.5.

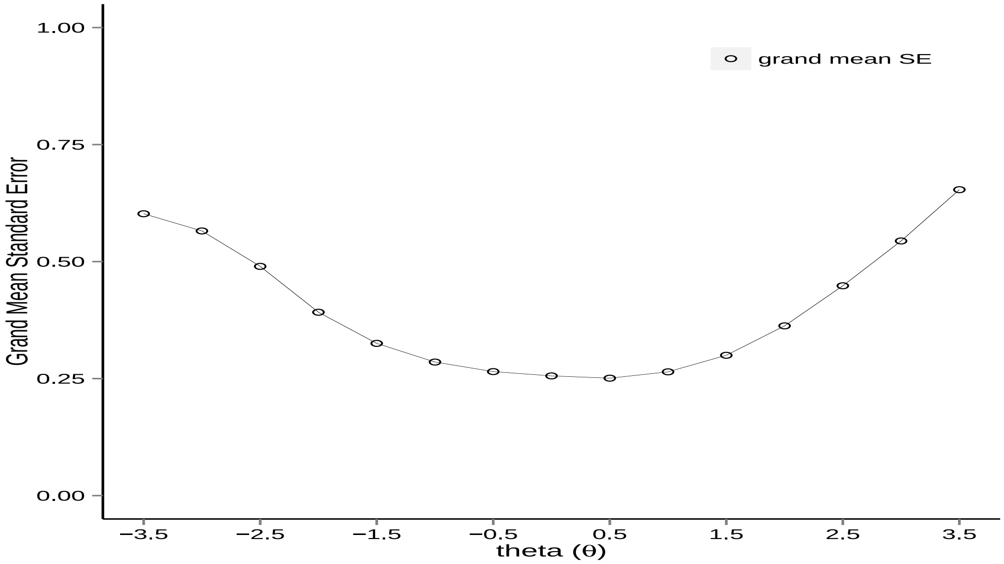

Figure 4.15: Conditional grand mean standard error averaged across all 32 study conditions.

Conditional standard errors for proportion condition

Grand mean conditional standard errors in terms of proportions are calculated to highlight the impact of the proportion of polytomous test units. Figure 4.16 presents the grand mean standard errors of MST with 40 total points for four levels of proportion of polytomous test units averaged across two routing module designs and two MST structures. Virtually no differences of measurement precision occurred in the medium and hard difficulty levels. In the easy difficulty levels, however, the MST with the 70% proportion of polytomous test units condition resulted in the largest standard errors, followed by the MST with the 50% condition.

Figure 4.17 shows the grand mean standard errors of MST with the 60

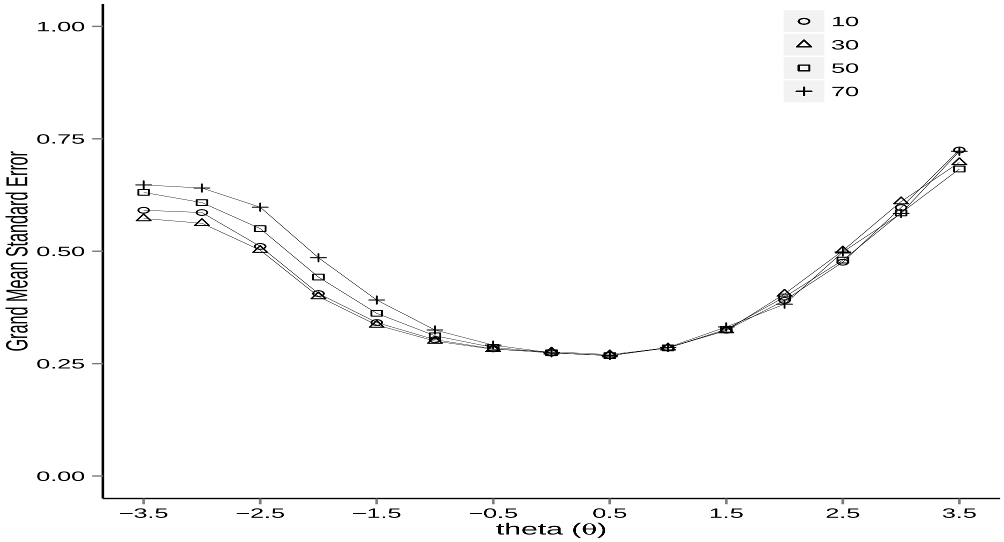

Figure 4.16: Conditional grand mean standard errors of MST with 40 total points for four levels of proportions

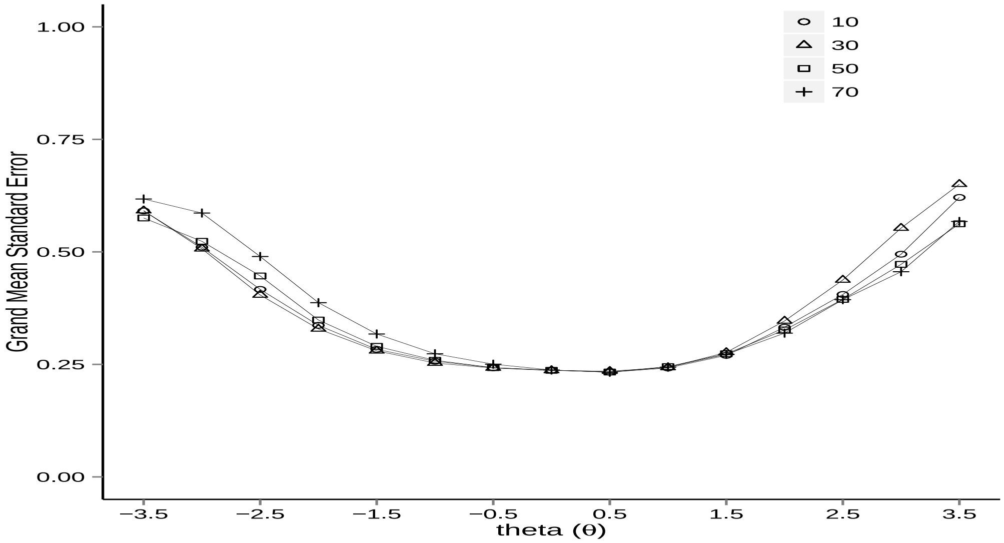

Figure 4.17: Conditional grand mean standard errors of MST with 60 total points for four levels of proportions total points for four levels of proportion of polytomous test units averaged across two routing module designs and two MST structures. All proportion conditions showed similar performance across the theta scale, while MST with the 70% proportion of polytomous test units showed the largest standard errors in low ability levels. In the hard difficulty levels, however, MST with the 30% proportion of polytomous test units showed the largest standard errors.

Conditional standard errors for total points and MST structure conditions

Figure 4.18 shows the grand mean conditional standard errors for MST with the 10% proportion of polytomous test units and dichotomous routing module design. MST with 60 total points resulted in smaller standard errors than with 40 total points, although the difference was smaller in the medium difficulty levels compared to the easy and hard difficulty levels. Compared to MST with the 122 structure, MST with the 133 structure tended to produce smaller standard errors; no differences emerged in the medium difficulty levels. When the total points and structure were considered together, an interaction between the two conditions was observed. While MST with the 60 total points condition produced smaller standard errors, MST with the 60 total points and 122 structure resulted in larger standard errors in hard difficulty levels than MST with the 40 total points and 133 structure.

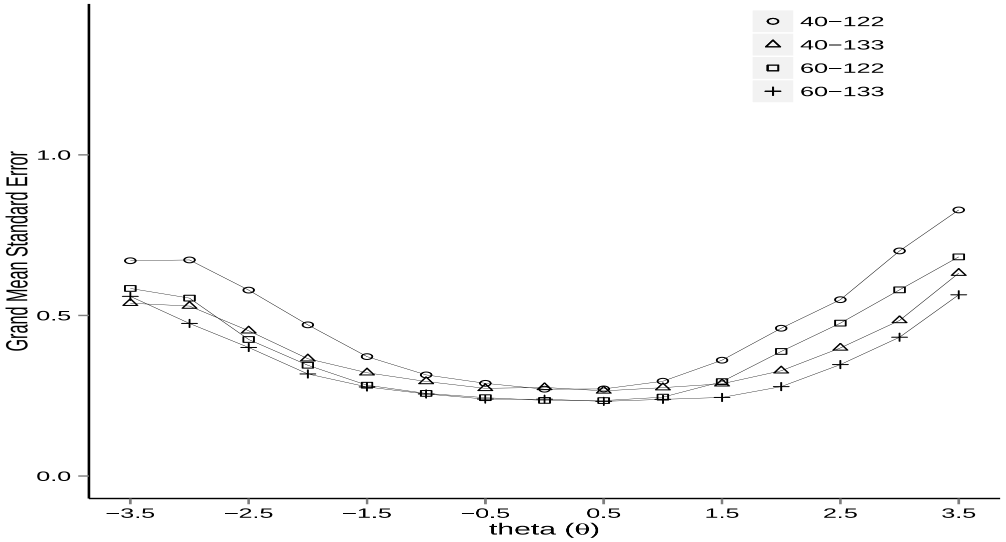

Figure 4.18: Conditional grand mean standard errors for 10% proportion of polytomous test units and dichotomous routing module condition

Figure 4.19 presents the grand mean conditional standard errors of MST with the 10% proportion of polytomous test units and mixed routing module. Similar patterns of U-shaped conditional standard error curves were observed. In addition, a similar interaction between total points and the MST structure in hard difficulty levels was observed (i.e., the higher measurement precision of MST with the 40 total points and 133 MST structure than MST with the 60 total points and 122 MST structure condition).

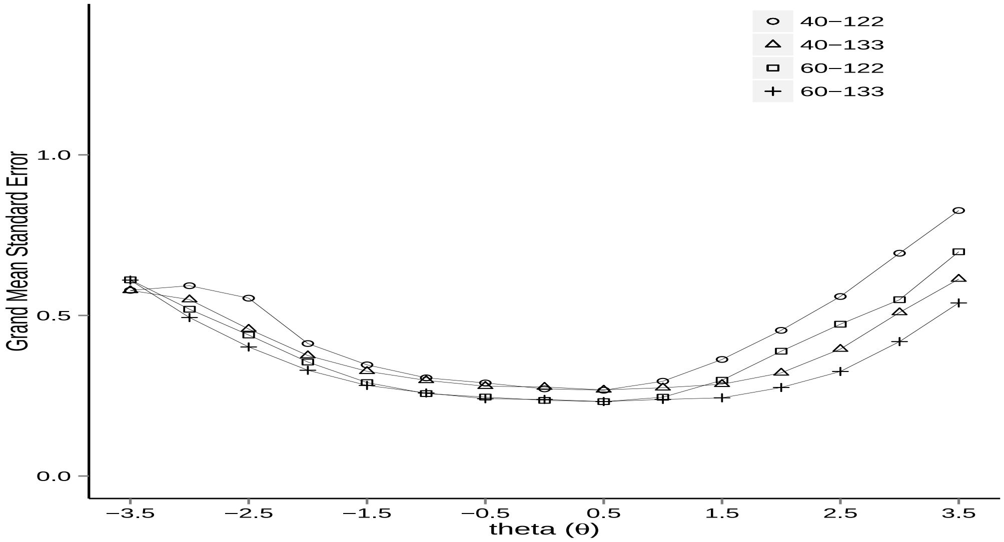

Figure 4.19: Conditional grand mean standard errors for 10% proportion of polytomous test units and mixed routing module condition

Figures 4.20 and 4.21 present the conditional grand mean standard errors of the 30% proportion of polytomous test units condition for the dichotomous and mixed routing modules, respectively. For both plots, MST with the 60 total points and 133 structure resulted in the highest measurement precision, where standard errors were kept below 0.5 across the wide range of ability scale. For both the 40 and 60 total point conditions, the MST with the 122 structure resulted in somewhat elevated standard errors in the hard difficulty levels.

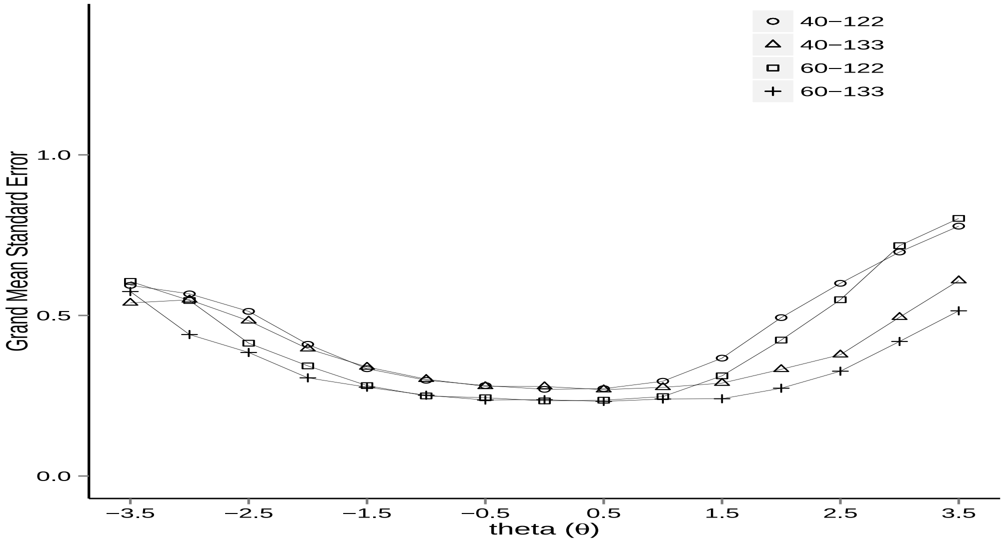

Figure 4.20: Conditional grand mean standard errors for 30% proportion of polytomous test units and dichotomous routing module condition

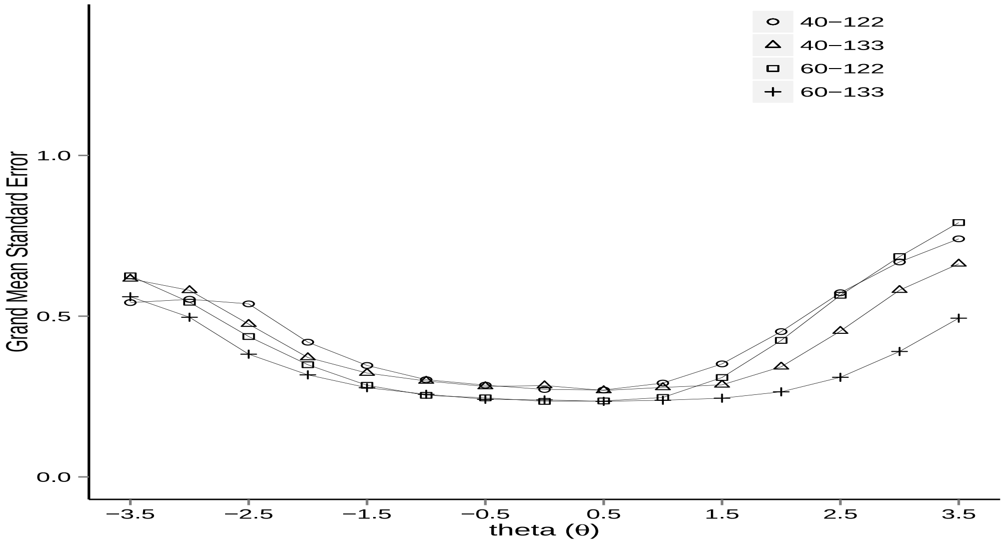

Figure 4.21: Conditional grand mean standard errors for 30% proportion of polytomous test units and mixed routing module condition

Figures 4.22 and 4.23 show the conditional grand mean standard errors of MST with the 50% proportion of polytomous test units for the dichotomous and mixed routing modules, respectively. MST with the 60 total points and 133 structure resulted in the highest measurement precision across ability levels, while MST with the 40 total points and 122 structure showed the largest conditional standard errors for wide range of ability levels. For the mixed routing module condition, the performance of MST with the 40 total points and 122 structure degraded rapidly as theta moved beyond 1.0.

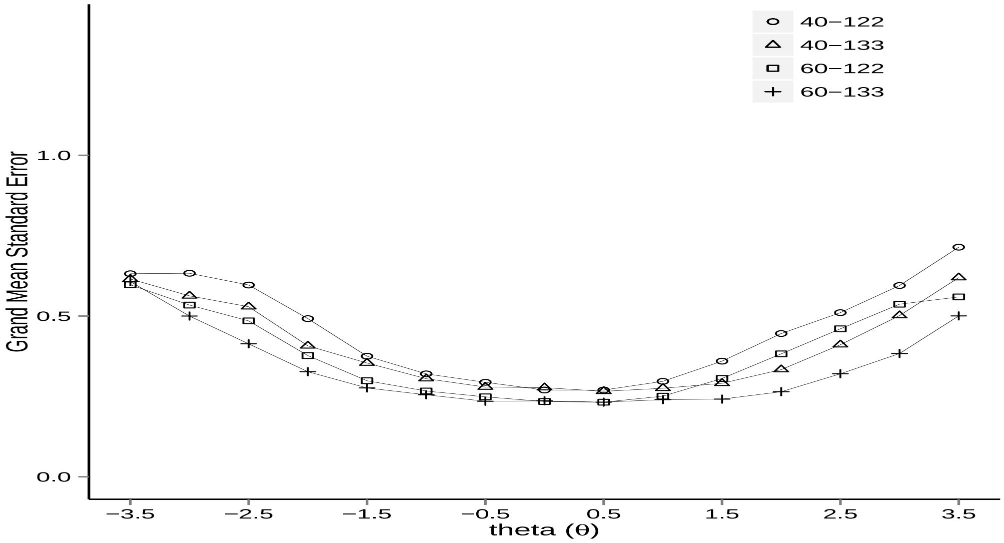

Figure 4.22: Conditional grand mean standard Errors for 50% proportion of polytomous test units and dichotomous routing module condition

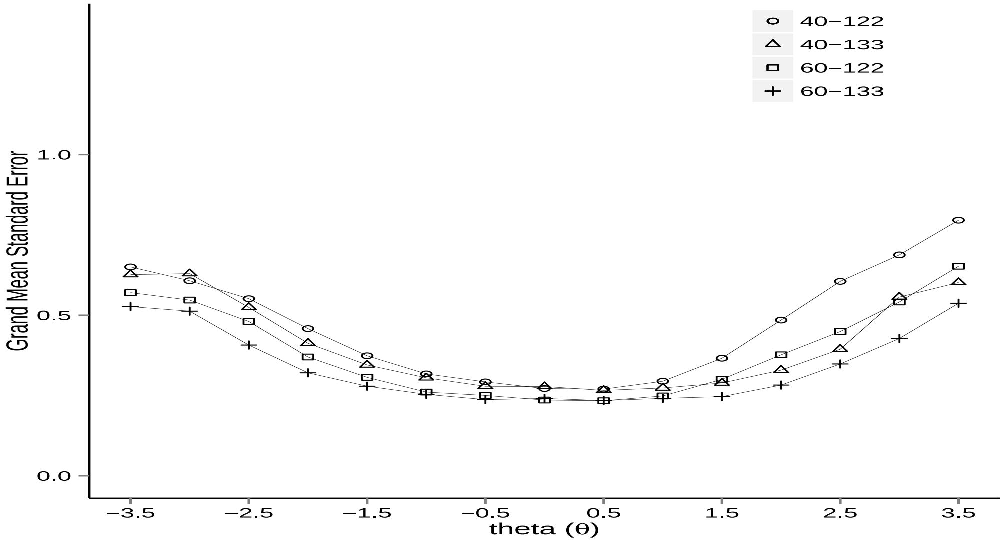

Figure 4.23: Conditional grand mean standard Errors for 50% proportion of polytomous test units and mixed routing module condition

Figures 4.24 and 4.25 present conditional grand mean standard errors of the MST with the 70% proportion of polytomous test units for the dichotomous and mixed routing module conditions, respectively. Standard errors in easy difficulty levels were larger than the 10%, 30%, and 50% proportion of polytomous test units conditions. MST with the 122 structure conditions, in particular, showed a rapid decline in measurement precision as theta moved away from the medium difficulty levels. For both routing module designs, MST with the 60 total points and 133 structure resulted in the smallest standard errors across all difficulty levels, while MST with the 40 total points and 122 structure produced the largest standard errors.

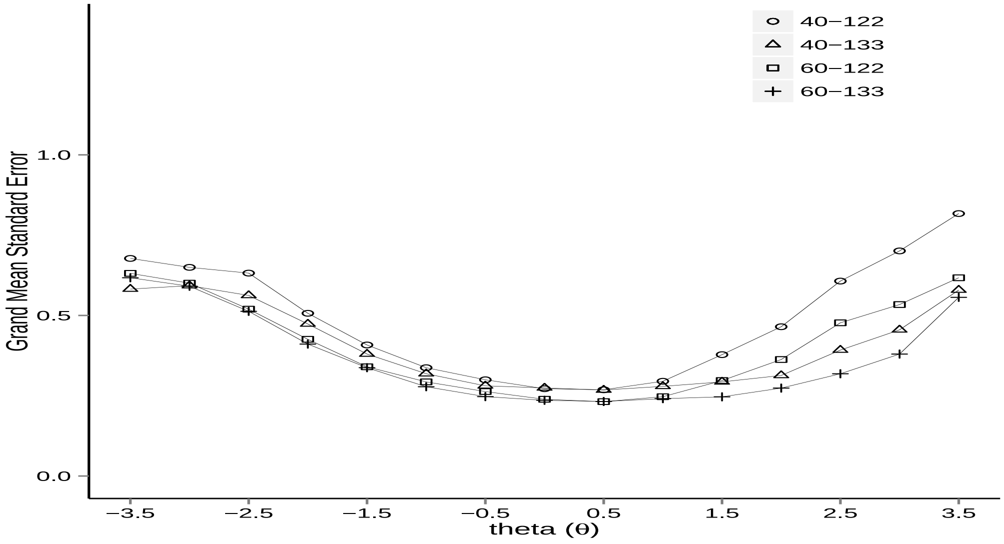

Figure 4.24: Conditional grand mean standard Errors for 70% proportion of polytomous test units and dichotomous routing module condition

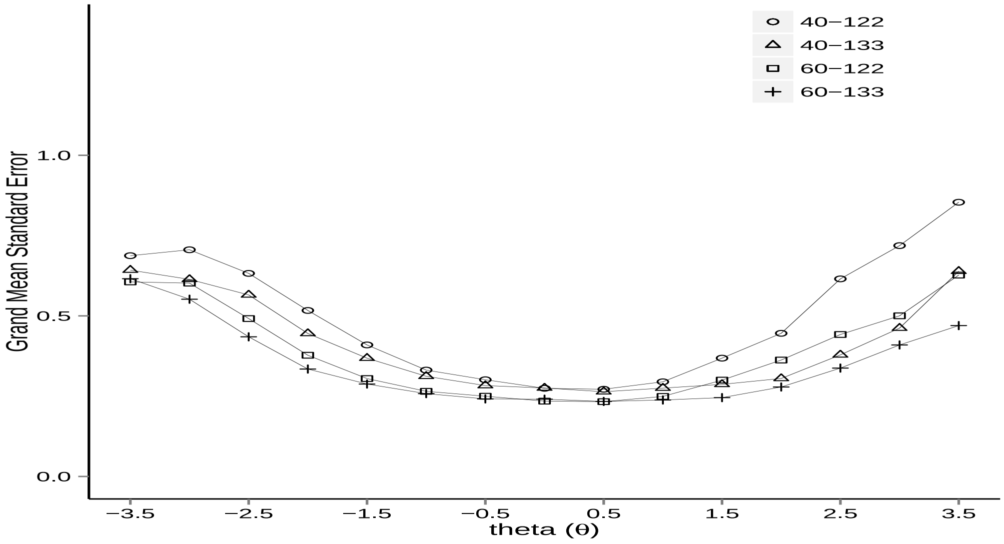

Figure 4.25: Conditional grand mean standard Errors for 70% proportion of polytomous test units and mixed routing module condition

Conditional standard errors across routing module design

Figure 4.26 depicts the conditional grand mean standard errors of MST of two routing module designs given the condition of 70% proportion of polytomous test units, 60 total points, and 133 MST structure. This is the only condition to show a difference between the two routing module designs considered in this study. The result shows that standard errors were larger for MST with the dichotomous routing module design for low ability levels than the mixed routing module design.

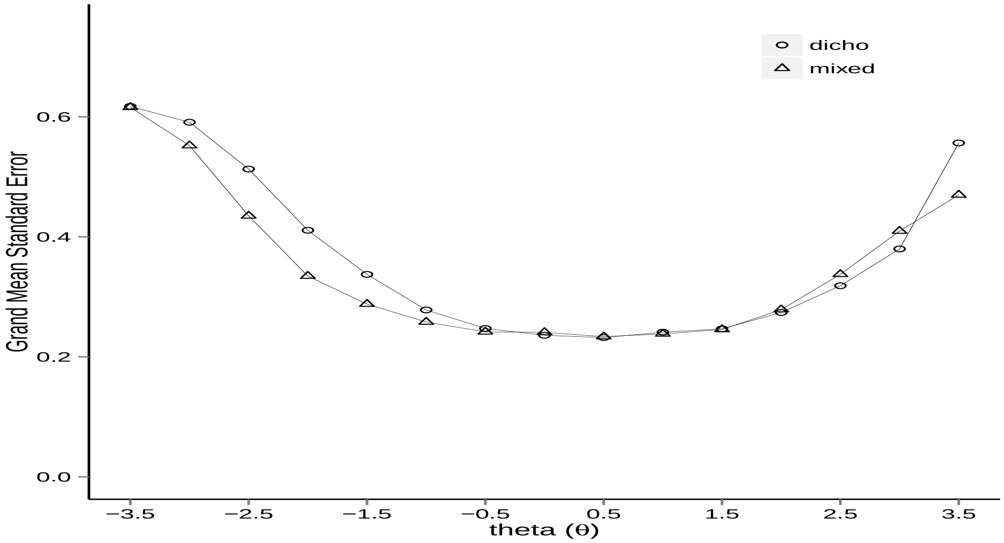

Figure 4.26: Conditional grand mean standard errors for 70 percent, 60 total points, and 133 MST structure

In summary, MST with 60 total points resulted in higher precision than MST with 40 total points condition (i.e., average 0.062 smaller standard error). However the performance advantage of the former over the latter is minimal in the medium difficulty levels (i.e., where theta is between -1.0 and 1.0). Compared to MST with the 122 structure, MST with the 133 structure resulted in smaller mean standard errors across a wide range of difficulty levels. The improved precision was more pronounced as the theta deviated away from the medium difficulty levels. In addition, an interaction between total points and MST structure was observed, in which MST with larger total points did not necessarily produce higher precision across difficulty levels. For example, MST with 40 total points and the 133 MST structure produced smaller standard errors than MST with 60 total points and the 122 MST structure in hard difficulty levels. In terms of proportion, MST with the 70% condition showed elevated standard errors in the easy difficulty levels compared to other proportion conditions. This result corresponds to the low MST test information presented in Figures 4.9 and 4.8.

The routing module design produced different results for MST with the 70% proportion of polytomous test units, 60 total points, and 133 structure. Standard errors of MST with the dichotomous routing module design were larger than MST with the mixed routing module design in easy difficulty levels.

Chapter 5

DISCUSSION

This study investigated the impacts of design variations of mixed-format MST on measurement precision. Four variables were manipulated in the study: the four levels of proportions of polytomous test units, two levels of total points, two levels of MST structures, and two levels of routing module designs, resulting in a total of 32 conditions. The simulation results included the descriptive statistics of estimated thetas and the correlation between known and estimated thetas. The measurement precision of MST on a wide range of ability levels was evaluated from grand mean standard errors conditional on θs, and MST test information functions were used to investigate the quality of MST constructions.

This chapter contains three sections discussing the study results. The first section discusses the research questions, study results, and findings. The second section focuses on the implications of findings for practical applications. Finally, the third section discusses the limitation of the study and future research directions.

Research questions

The following research questions are answered in this study:

1. How does test length (total points) impact measurement accuracy for mixed-format MST? The current study considered two total points conditions: 40 total points recommended for the diagnostic and licensure tests (Ho & Dodd, 2008) and 60 total points for high-stakes mixed-format tests (Jiao, 2003). The study results revealed that the MST with 60 total points produced smaller measurement errors and biases in terms of mean statistics; for the 40 and 60 total point conditions, the mean RMSEs averaged over other design conditions were 0.427 and 0.369 and the mean biases averaged over other study conditions were 0.006 and 0.001, respectively. The simulation results in this study are consistent with previous findings from studies varying the test length in MST (Chen, 2010; Jodoin et al., 2006; Kim et al., 2012).

The current study, however, revealed that the performance advantage of the MST with 60 total points is not constant over the wide range of θs compared to the MST with 40 total points. The simulation results produced the grand mean standard error difference of 0.036 between MST with the 40 and 60 total points at medium difficulty levels (i.e., thetas ranging between -1.0 and 1.0). The difference in standard errors between MST with 40 and 60 total points in the medium difficulty levels could be predicted from Equation 3.14. Using peak information of 15 and 20 for MST with 40 total points and 60 total points, respectively, the expected standard error difference is $1/\sqrt{15}$

- $1/\sqrt{20}$, which is 0.035.

From this small difference of grand mean standard error, the impact of total points to the measurement precision seems trivial in the medium difficulty levels, or where the peaks of information functions are located. However, an interaction between total points and the MST structure was observed. In the hard difficulty levels, MST with the 60 total points resulted in less accuracy than the 40 total points when MST with 60 total points was built in 122 structure and MST with 40 total points was constructed in 133 MST structure. This indicates that the 122 MST structure might not be able to provide the desired precision for a wide range of ability levels even with an increased total points (e.g., 60). On the other hand, the 133 MST structure maintained measurement accuracy in both extremes of the ability levels. Among others, MST with the 60 total points and 133 structure resulted in the highest measurement accuracy and MST with the 40 total points and 122 structure produced the lowest measurement precision.

2. What are the important design features when creating modules, stages, and panels for mixed-format MST?

The findings from the current study suggest that the correct specification of target test information functions and accurate construction are essential for the desired measurement precision, as has been reported in previous studies (Luecht, 2000; Luecht & Burgin, 2003; Luecht & Nungester, 1998). The specification of target test information functions largely depends on the purpose of the test and test unit pool characteristics

(Zenisky et al., 2010). The height of the peaks and locations should be determined based on the purpose of test and whether the precision is desired in a narrow range of ability levels (i.e., classification) or wide range of ability levels (i.e., ability estimation). In addition, the target test information needs to be determined such that test forms can be constructed using the pool.

The carefully constructed MST with the 122 MST structure could achieve equivalent precision as MST with the 133 structure in focused ability levels. As MST with the 122 structure requires a fewer number of modules within a panel, it increases the number of panels that can be constructed in a given test unit pool. This results in a higher level of test security and exposure control. On the other hand, MST with the 122 structure could not provide the desired precision in low and high ability levels, even with the increased total points of 60. Thus, the 133 structure should be recommended if the test goal is measuring the ability for a wide range of ability levels.

The evidence suggests the importance of the routing module design. Previous studies have focused on the impact of the various routing module lengths of MST designs (Chen, 2010; Macken-Ruiz, 2008), whereas the current study varied the proportion of test unit types while holding the routing module length conditions constant. The studys findings showed that a purely dichotomous routing module design had a negative impact on measurement precision given the large proportion of polytomous test units in MST panels. Therefore, the routing module design warrants extra caution when the proportion of polytomous test units is large in a mixed-format

MST test design. In addition, the high usage of major pathways may suggest that the second and the third stages could be merged to form 12 or 13 structures on behalf of 122 and 133 structures. The two stage designs are expected to reduce the burden of building extra modules and stages while maintaining equal or similar measurement precision compared to three stage structures. In terms of pathway frequency, nearly 90% of examinees stayed on the major pathways, indicating that once the difficulties are determined for each examinee after they finish the routing module, their difficulties did not change.

3. How do the various MST structures under the mixed-format context differ in terms of measurement precision

Overall, MST with the 133 structure performed better than MST with the 122 structure, producing smaller standard errors over a wider range of difficulty levels. MST with the 122 structure performed equally as well as MST with the 133 structure in the medium ability levels. As modules in the 133 structure can cover easy, medium, and hard difficulties, they can achieve greater precision in low and high ability levels as well as medium difficulty levels. Three modules in the 133 structure could maintain relatively sufficient information for a wide range of levels compared to the two modules in the MST with the 122 structure. As the 122 structure focused the precision in a small range of levels while utilizing fewer resources compared to the 133 structure, it has an advantage over the 133 structure for licensure tests or classification tests as reported by Zenisky (2004). On the other hand, when the goal of the test is measuring ability for a wide range of levels, the 122 structure might not provide the necessary measurement precision. The improved precision of MST with the 133 structure indicates that the accuracy of a test for a wide range can be achieved from modules constructed to function for a wide range of difficulties. This finding is consistent with previous research suggesting three modules per stage for the desirable measurement precision (Armstrong, Jones, Koppel, & Pashley, 2004).

4. Are there performance differences among the varied proportions for test unit types in the mixed-format MST?

The simulation results revealed no noticeable difference among 10%, 30%, and 50% proportion of polytomous test units conditions in terms of measurement precision. However, MST with the 70% proportion of polytomous test units showed somewhat elevated standard errors for low ability levels, while performance remained similar for medium and high ability levels. Furthermore, MST with the 70% proportion of polytomous test units condition produced a narrow MST test information function, resulting in larger standard errors in low ability levels. Three issues might have cause this result. First, when a large proportion of the test units in MST is polytomous test units, the actual number of test units in the MST is reduced because a small number of test units can account for the required total scores. This reduced number of test units might impose a difficulty in constructing MST panels to the specified target information. Second, the test unit pool contains a small proportion of polytomous test units. Thus, a small number of polytomous test units should be distributed to multiple MST panels. When the proportion of polytomous test units is large, it is difficult to form the desired target information with limited resources. Third, the test unit pool is positively skewed and there is relatively small information in low ability levels. This is directly related to the small MST test information and increased standard errors in low ability levels.

5. How does item type distribution at the routing module impact the measurement precision of mixed-format MST?

The routing module should include sufficient information to determine pathways for each of the examinees. The decision at the routing module seems more important than routing decisions in the rest of the stage based on the observation of pathway frequency in Table 4.3 Therefore, a large amount of information in the routing module is desirable for assigning highly informative modules to individual test takers. Kim et al. (2012) also reported on the importance of the routing module design for classification accuracy in the context of mixed-format MST. Their study results showed that higher classification accuracy is achieved by increasing the information of the routing module.

Overall, the routing module design did not produce a noticeable impact on the measurement precision for the mixed-format MST. However, the simulation results found an interaction between the routing module design and proportion conditions. When the proportion of polytomous test units is large and the routing module is purely dichotomously scored, the constructed

MST showed smaller test information and large standard errors. Under this condition, modules in the second and third stages in the MST panels should be constructed using an even larger proportion or (or completely) polytomous test units. In addition, the number of modules in the second and third stages (e.g., 6 in 133 design) grows rapidly as the number of panels increases. An automated test assembly tool could not find a set of test units to accurately represent the shapes of target test information functions (e.g., the peak values and locations) from the test unit pool. In other word, the issue of large proportion of polytomous test unit tended to be aggravated with purely dichotomous routing module condition.

Conclusions and practical applications

The current study investigated the impact of a mixed-format test unit pool for the measurement performance of MST. Four factors were varied in the study (i.e., four proportions of polytomous test units, two total scores, two MST structures, and two routing module designs), forming 32 total conditions. The target test information functions were specified for each total score and MST structure condition, and linear programming was used to construct MST panels to meet the specified target test information functions. For the MST simulation, 100 replications were performed and 1,000 normally distributed examinees were generated in each replication.

MST with small and medium proportions of polytomous test units (i.e., 10%, 30%, and 50%) performed in similar way. Holding the structure condition constant, MST with a larger total score produced greater precision due to the additional information within the test form. The greater precision of ability estimates could be achieved when large information exists (i.e., medium difficulty levels), and the standard errors could be as low as 0.25 for MST with 60 total scores. MST with the 133 MST structure showed smaller standard errors for a wide range of ability levels. However, the differences between the 122 and 133 structures were trivial for the medium ability levels. Thus, the advantage of MST with the 122 structure is that it can provide high performance in focused ranges of ability while utilizing fewer test units. When the goal of the test is measuring examinees' abilities on a wide range of θ scales, the 133 structure is more appropriate for performing well in a wide range of ability levels. In addition, the routing module design did not affect MST performance for the 10%, 30%, and 50% proportions of the polytomous test units.

When a large proportion of the polytomous test units (i.e., 70%) was used, MST produced larger standard errors in low ability levels for two possible reasons. First, a large proportion of polytomous test units actually requires fewer test units as the total scores can be satisfied using a small number of test units. From the perspective of optimal test construction, it might be more difficult for a small number of test units to form a test that satisfies the specified target information. Second, a limited number of polytomous test units are included in the item pool. By utilizing a large number of polytomous test units, MST construction quickly exhausts the limited resources in the test unit pool. Often the solutions for the optimal test construction were not feasible during the test construction, and MST panels did not meet the target test information functions. In addition, when the test construction needed to be constructed using limited test unit types, the characteristics of the pool itself directly affected the test, regardless of the target information test the developer specified. In low ability levels, the limited information available in the test unit pool resulted in limited MST test information because the test forms had to utilize almost all polytomous test units within the pool.

The impact of the routing module design was observed when MST was constructed using a large proportion of polytomous test units. MST with purely dichotomous test units resulted in larger standard errors compared to MST with the mixed routing module design in the low ability levels. Using the purely dichotomous routing module design and a 70% proportion of polytomous test units, modules in the second and third stages of MST needed to be constructed almost purely with polytomous test units. As the number of polytomous test units in the final MST forms needs to be multiplied by the number of panels, the dichotomous routing module design condition quickly exhausts the polytomous test units in the pool, resulting in less information in the low ability levels. The findings in the study should help test developers and administrators make better decisions when the design factor includes the proportion of test unit types. The lower measurement precision caused by the use of a large proportion of polytomous test units was noted and possible explanations have been provided.

The mixed-format MST is an attractive test format that renders a balanced measurement using different test unit types; as a result, it has received great attention (Kim & Dodd, 2014). For example, the National Center for Education Statistics (NCES) recently considered implementing the mixed-format MST in the NAEP test design (Oranje et al., 2014). The findings in this study offer practical recommendations for test developers. First, the importance of the test unit pool should be emphasized. A high-quality mixed-format test unit pool will allow the test developers to construct MST with high precision for a wide range of ability levels while satisfying various requirements (e.g., content, test taking time, test scoring effort). Second, design factors considered in this study will provide test developers with the necessary information to develop tests. For example, the interaction between the routing module design and the proportion of the polytomous test unit indicates that each design factor should not be considered separately, but rather warrants a more holistic approach. In addition, test developers should consider the potential impacts of factors that are not included in this study. For instance, the routing module design could be forced by the availability of automated scoring engines. Without automated scoring engines for constructed-response test units, the routing modules are often designed using purely dichotomous test units so that the ability estimation can be done on-the-fly.

Limitations and future research directions

As with all studies, this study has certain limitations. First, the MST construction outcomes might have been affected by the test unit pool characteristics. For example, relatively fewer test units functioned around the theta point of -1.0 than for the medium and hard difficulty levels. Therefore, easy modules were more challenging to construct than medium and hard modules. The lack of easy test units in the test unit pool was more severe for polytomous test units, and the constructions of MST with a large proportion of polytomous test units were even more challenging. Future studies could replicate the current study using a different test unit pool.

The current study results revealed that the large proportion of polytomous test units created the most stringent condition for MST construction when combined with a purely dichotomous routing module design. Other important design elements are test-taking time and economy of scoring. Polytomous test units (e.g., constructed-responses) require more time to administer and the scoring is more expensive and time-consuming compared to dichotomous test units. The added administration and scoring costs might not justify the utilization of large proportion of polytomous tests unit in mixed-format MST. The current study only focused on the measurement accuracy of mixed-format MST. The practical implementation of this test administration requires the consideration of other design elements, such as content balancing, test taking time and the effort involved in scoring. Finally, it is desirable to be able to score constructed-response test units automatically, so that the routing decisions can be made without interventions from human graders. To date, there is no universal solution for the automatic scoring engines for all test unit formats and test developers need to rely on either ad hoc approaches or the third-party solutions, if they are available. A full-fledged mixed-format MST, therefore, requires further studies on flexible automatic scoring methods for test unit types in the mixed-format MST.

REFERENCES

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

Bennett, R. E., Morley, M., & Quardt, D. (2000). Three response types for broadening the conception of mathematical problem solving in computerized tests. *Applied Psychological Measurement*, *24*(4), 294–309.

Bergstrom, B. A., & Lunz, M. E. (1999). CAT for certification and licensure. In F. Drasgow & J. B. Olsen (Eds.), *Innovations in computerized assessment* (pp. 67–91). Mahwah, NJ: Lawrence Erlbaum Associates.

Betz, N. E., & Weiss, D. J. (1974). *Simulation studies of two-stage ability testing* (Tech. Rep.). Psychometric Methods Program, Department of Psychology, University of Minnesota.

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

Bock, R. D., & Mislevy, R. J. (1982). Adaptive eap estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, *6*(4), 431–444.

Boekkooi-Timminga, E. (1990). The construction of parallel tests from irt-based item banks. *Journal of educational and behavioral statistics*, *15*(2), 129–145.

Breithaupt, K. J., Mills, C. N., & Melican, G. J. (2006). Facing the opportunities of the future. In D. Bartram & R. K. Hambleton (Eds.), *Computer-based testing and the internet* (p. 219-251). Chichester: John Wiley and Sons.

Burt, W., Kim, S. J., Davis, L. L., & Dodd, B. G. (2003). *A comparison of item exposure control procedures using a CAT system based on the generalized partial credit model.* Paper presented at annual meeting of the American Educational Research Association, Chicago.

Chen, L.-Y. (2010). *An investigation of the optimal test design for multi-stage test using the generalized partial credit model.* Unpublished doctoral dissertation, University of Texas at Austin.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* New York, NY: Holt, Rinehart, and Winston (New York).

Cronbach, L. J., & Gleser, G. C. (1965). *Psychological tests and personnel decisions.* Urbana, IL: University of Illinois Press.

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.

Dodd, B. G. (1984). *Attitude scaling: A comparison of the graded response and partial credit latent trait models.* Unpublished doctoral dissertation, University of Texas at Austin.

Dodd, B. G., De Ayala, R., & Koch, W. R. (1995). Computerized adaptive testing with polytomous items. *Applied Psychological Measurement*, *19*(1), 5–22.

Dodd, B. G., & Koch, W. R. (1987). Effects of variations in item step values on item and test information in the partial credit model. *Applied Psychological Measurement*, *11*(4), 371–384.

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

Ercikan, K., Sehwarz, R. D., Julian, M. W., Burket, G. R., Weber, M. M., & Link, V. (1998). Calibration and scoring of tests with multiple-choice and constructed-response item types. *Journal of Educational Measurement*, *35*(2), 137–154.

Grady, M., & Dodd, B. G. (2009). *A comparison of exposure control procedures for mixed-format adaptive tests.* Paper presented at annual meeting of the National Council on Measurement in Education, San Diego, CA.

Hagge, S. L., & Kolen, M. J. (2011). Equating mixed-format tests with format representative and non-representative common items. In W. Lee & M. J. Kolen (Eds.), *Mixed-format tests: Psychometric properties with a primary focus on equating (volume 1)* (pp. 95–136). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa.

Haladyna, T. (2004). *Developing and validating multiple-choice test items.* Hillsdale, NJ: Lawrence Erlbaum.

Hambleton, R. K., & Cook, L. L. (1977). Latent trait models and their use in the analysis of educational test data1, 2, 3. *Journal of educational* *measurement*, *14*(2), 75–96.

Hambleton, R. K., & Swaminathan, H. (1984). *Item response theory: Principles and applications* (Vol. 7). New York: Springer.

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.

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

Ho, T. H., & Dodd, B. G. (2008). *A comparison of item exposure constraints for mixed-format tests in computerized adaptive testing with the partial credit model.* Paper presented at the annual meeting of the American Educational Research Association, New York, NY.

Hogan, T. P., & Murphy, G. (2007). Recommendations for preparing and scoring constructed-response items: What the experts say. *Applied Measurement in Education*, *20*(4), 427–441.

Hutchinson, T. (1991). *Ability, partial information, guessing: Statistical modeling applied to multiple-choice tests.* Adelaide, Australia: Rumsby Scientific.

Ihaka, R., & Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of computational and graphical statistics*, *5*(3), 299–314.

Jiao, H. (2003). *The effects of model specification error in item response theory-based computerized classification test using sequential probability ratio test.* Unpublished doctoral dissertation, The Floria State University, Tallahassee.

Jodoin, M. G. (2003a). Measurement efficiency of innovative item formats in computer-based testing. *Journal of Educational Measurement*, *40*(1), 1–15.

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

Jodoin, M. G., Zenisky, A. L., & 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.

Kim, J. (2010). *A comparison of computer-based classification testing approaches using mixed-format tests with the generalized partial credit* *model.* Unpublished doctoral dissertation, University of Texas at Austin.

Kim, 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.

Kim, J., Chung, H., Park, R., & Dodd, B. G. (2013). A comparison of panel designs with routing methods in the multistage test with the partial credit model. *Behavior Research Methods*, 1–12.

Kim, J., & Dodd, B. G. (2010). *Comparing computer-based classification testing approaches using mixed-format tests with the generalized partial credit model.* Paper presented at the annual meeting of the National Council on Measurement in Education, May 1–3, in Denver, CO..

Kim, J., & Dodd, B. G. (2014). Mixed-format multistage tests: issues and methods. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing: Theory and applications* (pp. 55–67). Boca Raton, FL: CRC Press.

Kingsbury, G. G., & Weiss, D. J. (1983). A comparison of IRT-based adaptive mastery testing and a sequential mastery testing procedure. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 237–254). New York: Academic Press.

Kinsey, T. L. (2003). *A comparison of irt and rasch procedures in a mixed-item format test.* Unpublished doctoral dissertation, University of North Texas at Denton.

Koch, W. R. (1983). Likert scaling using the graded response latent trait model. *Applied Psychological Measurement*, *7*(1), 15–32.

Konis, K. (2011). lpsolveapi: R interface for lp_solve version 5.5. 2.0. *R package version*, *5*, 2–0.

Land, A. H., & Doig, A. G. (1960). An automatic method of solving discrete programming problems. *Econometrica: Journal of the Econometric Society*, *28*(3), 497–520.

Linn, R. L., Rock, D. A., & Cleary, T. A. (1969). The development and evaluation of several programmed testing methods. *Educational and Psychological Measurement*, *29*(1), 129–146.

Lord, F. M. (1952). *A theory of test scores.* New York: Psychometric Society.

Lord, F. M. (1963). Elementary models for measuring change. In

C. W. Harris (Ed.), *Problems in measuring change* (pp. 21–38). Madison: University of Wisconsin Press.

Lord, F. M. (1971). A theoretical study of two-stage testing. *Psychometrika*, *36*(3), 227–242.

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

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

Luecht, R. M. (2000). Implementing the computer-adaptive sequential testing (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, New Orleans, LA.

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

Luecht, R. M., & Burgin, W. (2003). *Test information targeting strategies for adaptive multistage testing designs*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Luecht, R. M., & Nungester, R. J. (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 W. J. van der Linden, R. M. Luecht, & C. A. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 117–128). Dordrcht, The Netherlands: Kluwer.

Macken-Ruiz, C. L. (2008). *A comparison of multi-stage and computerized adaptive tests based on the generalized partial credit model*. Unpublished doctoral dissertation, University of Texas at Austin.

Masters, G. N. (1982). A rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174.

Muraki, E. (1992). A generalized partial credit model: Application of an em algorithm. *Applied Psychological Measurement*, *16*(2), 159–176.

Muraki, E. (1993). Information functions of the generalized partial credit model. *Applied Psychological Measurement*, *17*(4), 351–363.

Oranje, A., Mazzeo, J., Xu, X., & Kulick, E. (2014). A multistage testing approach to group-score assessments. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing: Theory and applications* (pp. 371–390). Boca Raton, FL: CRC Press.

Ostini, R., & Nering, M. L. (2006). *Polytomous item response theory models.* Thousand Oaks, CA: Sage Publication.

Park, R., Kim, J., Chung, H., & Dodd, B. G. (2011). *Linear programming modeling using the derivative objective function for automated test assembly.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Park, R., Kim, J., Chung, H., & Dodd, B. G. (2012). *Enhancing pool utilization in constructing the multistage test using mixed-format tests.* Paper presented at the annual meeting of the National Council on Measurement in Education, April 12–16, in Vancouver, Canada.

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

Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (2002). Issues in innovative item types. In C. G. Parshall, J. A. Spray, J. C. Kalohn, & T. Davey (Eds.), *Practical considerations in computer-based testing* (pp. 70–91). New York: Springer.

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

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 meeting of the National Council on Measurement in Education, Montreal, Quebec.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests* (Vol. 1). Chicago: University of Chicago Press.

Reshetar, R., & Melican, G. J. (2010). *Design and Evaluation of Mixed-Format Large Scale Assessments for the Advanced Placement Program (AP).* Paper presented at the annual meeting of the American Education Research Association in Denver, CO.

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

Rosa, K., Swygert, K. A., Nelson, L., & Thissen, D. (2001). Item response theory applied to combinations of multiple-choice and constructed-response items-scale scores for patterns of summed scores. In D. Thissen & H. Wainer (Eds.), (pp. 253–292). Mahwah, NJ: Lawrence Erlbaum.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, *34*(17).

Theunissen, T. (1986). Some applications of optimization algorithms in test design and adaptive testing. *Applied Psychological Measurement*, *10*(4), 381–389.

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, *51*(4), 567–577.

Thissen, D., Steinberg, L., & Mooney, J. A. (1989). Trace lines for testlets: A use of multiple-categorical-response models. *Journal of Educational Measurement*, *26*(3), 247–260.

van der Linden, W. J. (1987). Automated test construction using minimax programming. In W. J. van der Linden (Ed.), *IRT-based test construction* (pp. 1–16). The Netherlands: Department of Education, University of Twente Enschede.

van der Linden, W. J. (1998). Optimal assembly of psychological and educational tests. *Applied Psychological Measurement*, *22*(3), 195–211.

van der Linden, W. J. (2005). *Linear models for optimal test design.* New York: Springer.

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

Wainer, H. (1995). Precision and differential item functioning on a testlet-based test: The 1991 law school admissions test as an example. *Applied Measurement in Education*, *8*(2), 157–86.

Wainer, H., & Mislevy, R. J. (1990). Item response theory, item calibration, and proficiency estimation. In H. Wainer et al. (Eds.), (pp. 233–272). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Wainer, H., & Thissen, D. (1987). Estimating ability with the wrong model. *Journal of Educational and Behavioral Statistics*, *12*(4), 339–368.

Wainer, H., & Thissen, D. (1993). Combining multiple-choice and constructed-response test scores: Toward a marxist theory of test construction. *Applied Measurement in Education*, *6*(2), 103–118.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement*, *6*(4), 473–492.

Xing, D. (2001). *Impact of several computer-based testing variables on the psychometric properties of credentialing examinations*. Unpublished doctoral dissertation, University of Massachusetts, Amherst.

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.

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 at 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. Glas (Eds.), *Elements of adaptive testing* (pp. 355–372). New York: Springer.

Zenisky, A. L., & Sireci, S. G. (2002). Technological innovations in large-scale assessment. *Applied Measurement in Education*, *15*(4), 337–362.