# A TOP-DOWN APPROACH FOR OPTIMALLY DESIGNING MULTISTAGE-ADAPTIVE TESTS

HWANGGYU LIM

# A TOP-DOWN APPROACH FOR OPTIMALLY DESIGNING MULTISTAGE-ADAPTIVE TESTS

A Dissertation Presented

by

HWANGGYU LIM

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

DOCTOR OF PHILOSOPHY

May 2019

College of Education

Research, Educational Measurement, and Psychometrics

# A TOP-DOWN APPROACH FOR OPTIMALLY DESINING MULTISTAGE-ADAPTIVE TESTS

A Dissertation Presented

by

HWANGGYU LIM

Approved as to style and content by:

Craig S. Wells, Chair

April L. Zenisky, Member

Stephen G. Sireci, Member

Timothy C. Davey, Member

Jennifer Randall
Associate Dean of Academic Affairs
College of Education

# DEDICATION

To my beautiful and loving Jain.

# ACKNOWLEDGMENTS

I am grateful to my committee, family, and colleagues who so generously contributed to my study. First, I give my deep thanks to the chair of my dissertation committee, my professional mentor, and my supervisor through my graduate school, Professor Craig S. Wells, not only for his tremendous academic support, but also for giving me so many wonderful opportunities. Without his support and encouragement, I would have not finished my dissertation.

I am also deeply thankful to my academic mentor, Professor Ronald K. Hambleton for his instruction, support, and enthusiasm for research. For my whole school life, he always shared his valuable time to teach me, give me advice for research, and provide me with considerable knowledge. Similar, my special thanks goes to Professor Stephen G. Sireci who has been a truly dedicated mentor. He has always encouraged me to work and study better throughout my years as a graduate student.

I am also hugely appreciative to Professor April L. Zenisky, especially for sharing her expertise so willingly for my dissertation. Profound gratitude goes to Dr. Tim Davey who was my great mentor during the Summer internship program at Educational Testing Service (ETS). Due to his unselfish help, insight, and support, I could extend the research project at ETS to my dissertation.

Special mention goes to a number of former and current students from REMP for helping me to complete this study. I am especially grateful for my friendship with Hyunjoo Jung, as she always encouraged me and shared her academic expertise with me.

Love and gratitude to Jain, my wife, for her unbelievable encouragement and support throughout my graduate school. She is the most important person in my world and I dedicate this thesis to her.

# ABSTRACT

# A TOP-DOWN APPROACH FOR OPTIMALLY DESIGNING MULTISTAGE-ADAPTIVE TESTS

MAY 2019

HWANGGYU LIM, B.A., YONSEI UNIVERSITY

M.A., YONSEI UNIVERSITY

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Craig S. Wells

In multistage-adaptive testing (MST), there are many interrelated design variables that impact the nature and quality of ability estimation. Previous research has identified general principles for the effective design of MSTs in terms of measurement performance. However, those principles are unlikely to apply uniformly to every testing context.

The purpose of this dissertation is to propose a process of finding an MST design that has optimal measurement properties, given a specific set of test circumstances. To achieve this goal, an efficient strategy was introduced at each of three phases to discover the optimal design of the MST; constructing MSTs, systematically searching a design space of the MST, and evaluating the MST performance. For the first phase, a top-down approach was applied in this study. For the second phase, a way to systematically search the parameterized design space of an MST was used. For the third phase, a new analytical evaluation method for MST was proposed.

In the dissertation, Study 1 proposed a new analytical evaluation method for MST. Using this new approach, measurement precision of ability estimation and classification accuracy could be derived analytically. The simulation results indicated that the new analytical method produced more exact measurement properties of an MST than the Monte Carlo simulation method. Therefore, the new analytical method would be the most efficient and competitive tool to asses measurement performance of an MST among other evaluation methods.

Study 2 proposed a process to find a design of an MST that shows optimal measurement properties applying the three efficient strategies, given a specific set of testing context. The process consists of four important features: (1) setting a testing circumstance and MST design space, (2) systematically searching the MST design space using the top-down approach, (3) analytically evaluating measurement performance of an MST, and (4) computing objective functions. The suggested process was applied to a real item pool from a large-scale assessment. The results of the application study provided evidence that the process could be generalized to more complex and realistic test circumstances to create optimal designs of MST.

## TABLE OF CONTENTS

Page

CHAPTER

## LIST OF TABLES

Table Page

## LIST OF FIGURES

Figure Page

# CHAPTER 1

# INTRODUCTION

## 1.1 Background

A multistage-adaptive test (MST) uses a specific adaptive test design that tailors test difficulty to the performance level of an individual examinee. In recent years, MSTs have become increasingly popular as an alternative to conventional linear tests and item-level computerized adaptive tests (CATs). For example, many operational testing programs have replaced the paper-and-pencil linear test or the CAT with MST (e.g., American Institute of Certified Public Accountants (AICPA) Examination, National Council Licensure Examination (NCLEX), and Graduate Record Examination (GRE)). The primary reason for the popularity of MSTs is that they provide a balanced compromise between the linear test and the CAT (Hendrickson, 2007). Because the MST is an adaptive test, it is more efficient and precise in estimating an examinee proficiency compared to a linear test in which all examinees respond to test forms that are not tailored to each examinee's proficiency (Jodoin, Zenisky, & Hambleton, 2006; Kim & Plake, 1993). Although MSTs are less efficient than CATs, it is known that measurement precision of MST is still quite comparable to the CAT when the test is carefully designed (Luecht & Nungester, 1998; Xing & Hambleton, 2004).

An MST possesses several practical advantages that make the MST a favorable choice over CAT in an operational testing program (Melican, Breithaupt, & Zhang, 2010; Stark & Chernyshenko, 2006). First, because the tests are pre-assembled, the MST allows for subject matter experts and other stakeholders to review the psychometric and content properties of tests prior to publication. This property of MST is not only desirable for quality-control, but also it enables test developers to satisfy more complex and sophisticated content specifications in an MST since sometimes certain content requirements are difficult to quantify in the automated test assembly (ATA) process (Stark & Chernyshenko, 2006). Second, examinees are able to skip test items, revise their responses, and return to previous items for reviewing them within a stage, while CAT prohibits examinees from reviewing and skipping items. Therefore, MST could provide more comfortable testing circumstances to test-takers. Third, MST requires less computing power than CAT for ability estimation and item selection because MST only needs to compute interim proficiency estimates after each set of items instead of after each item as in CAT (Han & Guo, 2014).

An MST has special terminologies in terms of its design. In an MST, a test administration unit is called a *panel*, which is a group of pre-assembled item sets called *modules*. The MST panel is divided into several *stages* and each stage in the panel consists of multiple modules. Modules within the same stage usually have different difficulty levels targeted to particular levels of proficiency. During the process of testing, the combination of modules across stages that an examinee is administered to finish the test is called a *route* or *pathway*. In this study, both the route and pathway are used interchangeably.

Figure 1 illustrates two configurations of MST panels: 1-3 (left panel) and 1-3-3 (right panel). In Figure 1, E, M, and H stand for easy, moderate (or medium), and hard difficulty levels, respectively. Note that there are seven routes in the 1-3-3 MST of Figure 1 because two routes of 1M-2E-3H and 1M-2H-3E are removed from the panel.

Removing certain routes is a possible strategy in operational testing to prevent capricious proficiency changes due to cheating of items or brain dump (Luo & Kim, 2018).

In an MST panel, there is usually only one module at the first stage and it is called a routing module or router. Among all possible routes in a panel, there are special pathways called primary routes where subsequent modules after the first stage have the same difficulty level. For example, there are three primary routes in the 1-3-3 MST in Figure 1. In that MST panel, low proficiency examinees are likely to take the 1M-2E-3E, moderate proficiency examinees would tend to take the 1M-2M-3M, and high proficiency examinees are likely to be administered the 1M-2H-3H. Previous studies have shown that a large proportion of examinees (i.e., approximately more than 70%) are given the primary routes while taking the exam (Kim, Chung, Park, & Dodd, 2013; Luo & Kim, 2018; Zenisky, 2004). Other than primary routes, the rest of pathways such as the 1M-2E-3M and 1M-2H-3M in the 1-3-3 MST are called secondary or ancillary routes.

## 1.2 Statement of Problem and Its Significance

When implementing an MST, there are many interrelated design variables that impact the nature and quality of ability estimation (e.g., number of stages, number of distinct difficulty levels at each stage, module and test length, cut scores for routing examinees to next the modules, scoring methods, a population of examinees, item bank, and content requirements). Therefore, the process of test development involves a series of critical decisions to design an MST for the intended purpose of the test (Zenisky, Hambleton, & Luecht, 2010).

Constructing an MST is very flexible because test developers can customize the design of an MST as they want due to the existence of many design factors, which is one of the advantages in MST. However, this also implies that it does not seem feasible to discover a truly optimal design that shows the best measurement properties fitted to every testing circumstance. Although previous studies have documented the quality of measurement varying design variables in MST and identified general principles for the effective design of MSTs in terms of measurement performance (e.g., Hambleton & Xing, 2006; Luecht & Burgin, 2003; Luo & Kim, 2018; Park, Kim, Chung, & Dodd, 2014; Wang, Fluegge, & Luecht, 2012), those principles are unlikely to apply uniformly to every testing context. Rather, an MST design that is optimal in some sense for one testing program may work poorly for another. The number and nature of the items available, the rigor of the content requirements, and the location and scale of the examinee proficiency distribution are all factors that dictate whether a given MST design will work well or poorly under a specific testing situation. Therefore, it is necessary to develop an algorithm and process for finding, given a concrete set of test circumstances, a specific MST design that is optimal in some sense.

Because a theoretical space of design variables of MST is enormous, it is not easy to evaluate measurement properties of all possible combinations of MST design variables. One practical solution is to restrict a scope of design factors so that the combination of design variables has a reasonably limited range, depending on a testing context. Even with the restricted range, however, there may be still too many combinations to assess their measurement properties. To deal with this problem, more efficient strategies are necessary at three phases when finding an optimal MST design: (1) assembling MST given a certain set of design variables, (2) searching many combinations of design variables, and (3) evaluating the measurement performance of

MSTs. With more efficient strategies at each of those phases, a more broad range of MST design variables could be searched effectively and the performance of the MST would be assessed more quickly.

### 1.3 Purpose of Study

The purpose of this dissertation is to propose a process of finding an MST design that has optimal measurement properties in some sense, given a specific set of test circumstances. To achieve this goal, an efficient strategy is introduced at each of three phases, which are constructing MSTs, searching design variables of MST, and evaluating the MST performance, to discover the optimal design of an MST.

For the assembly of an MST, a top-down approach is applied in this study. In MST, a test is built using either a top-down assembly method or a bottom-up assembly method. The bottom up approach is a “divide-and-conquer” method because a test level specification for the statistical targets, content, and other features is divided into the module level and the modues are mixed-and-matched whereas the top-down strategy requires only test level specifications for the statistical targets and other non-statistical constraints to build a test (Luecht & Nungester, 1998). The reason of using the top-down approach in this study is that the computer algorithm of automated test assembly (ATA) identifies an optimal partition of test-level design variables into modules as well as achieves the optimal measurement precision (Luo & Kim, 2018). Thus, the top-down approach simplifies the design process compared to the bottom-up approach.

For the second phase, a way to systematically search the parameterized design space of an MST is used. Especially, this strategy involves systematically varying targeted subpopulations of routes and iteratively applying the ATA process based on the top-down approach. For the third phase, a new analytical evaluation method for MST is proposed in the study, which is based on the equated number-correct (NC) scoring method (Stocking, 1996). In most MST studies, Monte Carlo (MC) based simulation methods have been used to evaluate the performance of an MST (e.g., Armstrong, Jones, Koppel, & Pashley, 2004; Hambleton & Xing, 2006; Jodoin et al., 2006; Luecht, 2003; Luo & Kim, 2018; Wang, 2017; Wang et al., 2012; Weissman, Belov, & Armstrong, 2007; Zenisky, 2004). These simulation studies usually require a lot of time and effort to set up and conduct a simulation. Since measurement precision (e.g., the conditional standard error of ability estimates) of an MST is computed analytically with the new method, however, the evaluation is more exact and faster than those based on a simulation, which is important advantage when assessing a large number of design factors.

This dissertation consists of two studies. Because the analytical evaluation of MST performance is a new approach, it is necessary to provide evidence that the method works well. Accordingly, Study 1 introduces the new analytic method and demonstrates that the proposed method predicts measurement properties of an MST accurately. Then, Study 2 proposes a procedure to find a design of an MST, given a specific set of testing circumstances, that shows optimal measurement properties by means of the three efficient strategies.

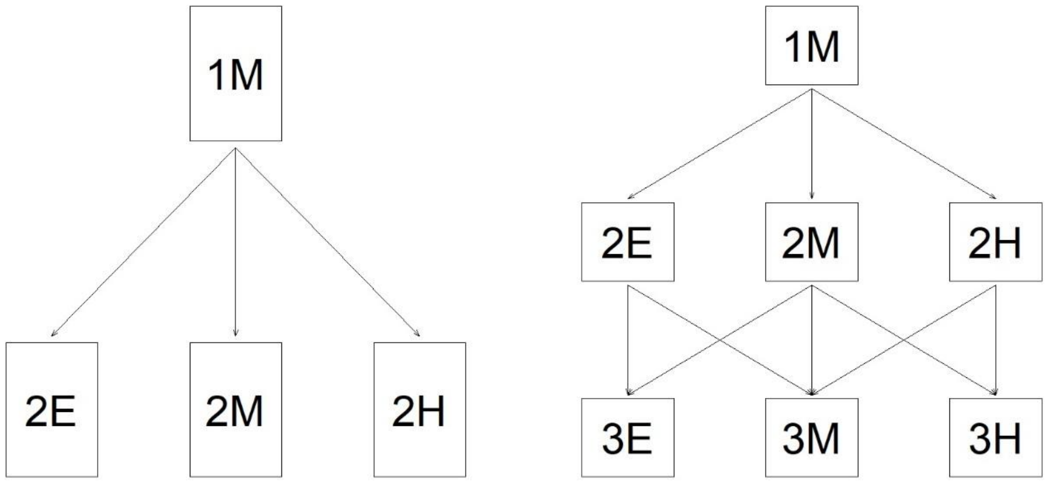

**Figure 1.** Examples of a 1-3 MST panel (left) and 1-3-3 MST panel (right)

# CHAPTER 2

# LITERATURE REVIEW

This study introduces a new analytical evaluation method of MST performance and proposes a process to discover an MST design that has optimal measurement properties using the analytical evaluation method given a specific testing context. Therefore, it is necessary to overview theories behind MST to give strong background to this study. The overview especially focuses on the basic design variables of MST, test assembly methods, evaluation methods of MST, and scoring and routing methods. Thus, this chapter consists of four sections and highlights research and practice related to an optimal design of MST. The first section reviews important design considerations in the development of MST. The second section discusses two evaluation methods of MST which are simulation-based and analytical methods. The third section provides some practical issues related to the test assembly of MST. The fourth section deals with some background of scoring and routing methods in MST.

## 2.1 MST Design Considerations

Implementing an MST is a complex process due to the large number of highly interdependent design variables that significantly affect the nature and quality of proficiency estimates. Test developers need to make a series of critical decisions based on various requirements of the intended purpose and expected consequences of a testing program in test development and administration (Zenisky et al., 2010). In fact, the existence of many design variables is one of the advantages in MST in that test developers can customize the design of an MST in numerous ways according to the testing program’s goal and purposes (Zenisky, 2004). Although it is not feasible to find one best MST design that fits every testing context in terms of optimal measurement performance, many studies have investigated the impact of varying design variables on the testing results and tried to find a reasonable combination of design variables that produce acceptable measurement precision under a particular testing context (Hambleton & Xing, 2006; Luecht & Burgin, 2003; Luo & Kim, 2018; Park et al.,  2014; Wang et al., 2012).

Lord (1980, p. 129) provided an outline of several important design factors to consider when building a two-stage testing design in terms of measurement precision. Zenisky et al. (2010, p. 357) genealized Lord’s ideas to an *n*-stage MST with additional considerations that have been examined in previous MST research studies. Among the MST decision variables described in Zenisky et al. (2010), the frequently investigated variables in MST research are as follows:

(1) number of stages;

(2) number of difficulty-level modules per each stage;

(3) total number of items in the test;

(4) number of items per each module;

(5) statistical characteristic of modules (i.e., shape of module information function);

(6) cut-points or methods for routing examinees to modules; and

(7) method for scoring stages and each *n*th-stage test.

In addition, Luecht and Burgin (2003) and Luecht (2014) decribed several MST panel design consderations which correspond to the 1-5 among the list above. Also, Wang et al. (2012) conducted an exhaustive comparative study to examine the accracy and efficiency of MST under various panel design conditions. Wang et al. (2012) not only addressed design variables such as the number of stages, the number of difficulty-level modules within each stage, and the number of items per module, but they also explored the interaction between MST panel design variables and the item bank size and/or item quality in the bank.

Since theoretical and practical concept of design variables of MST have an enormous range, addressing all of MST design variables is beyond the scope of this study. To find an optimal MST design with optimal measurement properties in some sense given a specific testing context, this study manipulates the characteristics of MST related to the design of panel configuration. Therefore, this section discusses the design considerations of MST, paying special attention to variables pertaining to MST panel design. To facilitate discussion of the MST panel design issues, each of these considerations are loosely clustered as related to either (1) shape of panel structure (e.g., 1-2-2 and 1-3-5 designs), (2) test length, (3) characteristics of module, or (4) item bank and examinee population.

### 2.1.1 Shape of Panel Structure

The shape of a panel structure for MST consists of a combination of the modules and stages and when considered together indicate the possible routes that an examinee could take while being administered the MST. In particular, the number of stages and the number of difficulty-level modules per each stage are primary concerns in designing an MST panel structure. Theoretically, the range of possible forms of the panel structure for an MST is innumerable and a large number of panel structures have been studied and/or used in practice. Several representative examples of panel structures in the majority of MST literature follow a 1-3 MST (e.g., Kim & Plake, 1993; Luecht & Nungester, 1998; Luo & Kim, 2018; Schnipke & Reese, 1997; Xing & Hambleton, 2004), 1-2-2 MST (e.g., Breithaupt & Hare, 2007; Park, Kim, Chung, & Dodd, 2017; Zenisky, 2004), 1-3-3 MST (e.g., Hambleton & Xing, 2006; Jodoin et al, 2006; Luo & Kim, 2018; Park et al., 2014), and 1-2-3-4 MST (e.g., Wang et al., 2012; Zheng, Nozawa, Gao, & Chang, 2012). For the number of stages, many MST research and applications have used two, three, and four stages. For the number of modules per stage, one module is frequently used at the first stage and the number of different-level modules increases across subsequent stages assuming one panel is built.

Since an MST is an adaptive test, the use of more stages in a panel, and more difficulty-levels of modules within the stages, allows for greater adaptation and more flexibility (Hendrickson, 2007). In the context of achievement tests, adding more stages and more modules per stage usually aims for more measurement precision in the tail areas of the ability scale. However, designing panels with more stages and modules also complicates the test assembly without necessarily adding more measurement precision and may result in the decrease of the overall quality of the MST (Hendrickson, 2007; Luecht, 2014; Yan, Lewis, & von Davier, 2014). For example, as more stages are used in a panel of MST, more potential pathways should satisfy the same statistical and nonstatistical constraints. As more different difficulty levels of modules are added within a stage, it might require extremely easy or difficult items for the modules that represent the extreme levels of ability, resulting in not meeting the target specification in the test assembly.

In fact, if test length is long-enough to obtain a high level of test information, not much difference is found between different panel structure designs when it comes to measurement efficiency and precision (Jodoin et al., 2006; Kim & Plake, 1993; Luo & Kim, 2018; Schnipke & Reese, 1997; Wang et al., 2012; Zenisky, 2004; Zheng et al., 2012). Under the two stage testing context, Kim and Plake (1993) indicated that the increase in the number of modules at the second-stage did not make any significant improvement of measurement accuary.

Jodoin et al. (2006) investigated the measurement properties of two panel designs of MST for a large-volume credentialing exam – a 40 item two-stage test and a 60 item three-stage test. In the study, the 60 item three-stage tests consistently produced strong psychometric properies such as more accurate abiltiy estimates, decision consistency, and decision accuracy than the 40 item two-stage tests. But, they observed that the 40 item two-stage tests performed nealy as well as the 60 item three-stage tests, claiming that including more stages does not significantly increase measurement precision.

Zheng et al. (2012) compared different panel designs of MST given a large-scale classification testing context. Their study used a three-stage design (1-2-4 MST) and a four-stage design (1-2-3-4 MST) with total length of 21 items and compared correct classification rates (CCR) for both panel designs. The results showed that the four-stage panel resulted in slightly higher CCR than the three-stage panel when the item pool was optimized and the overap of items within a stage was allowed. However, no consistent advantages of CCR in the four-stage panel were found.

Recently, Wang et al. (2012) found that with the regular item bank, MST panel designs with fewer modules within a stage (e.g., 1-2, 1-2-2 MSTs) were more effective in terms of achieving appropriate adaptation of the module difficulties than panel designs with more modules within a stage (e.g., 1-3, 1-3-3 MSTs). However, no substantial differences were observed between different MST panel designs in Luo and Kim (2018). In their study, the more compex panel designs (e.g., 1-2-2 and 1-3-3 MSTs) did not produce significantly increased measurement precision than the simpler panel design (e.g., 1-3 MST).

As Luecht (2014) noted, most previous literature indicates that designing an MST with more than three stages may be sufficient to produce an acceptable level of measurement precision as long as the designed MST provides adequate test lengths, degree of adaptation, and accumulation of measurement information to match the score precision and/or decision accuracy are provided. Also, researchers have shown that a maximum of four modules is desirable at any stage in general (Armstrong et al., 2004; Hendrickson, 2007).

### 2.1.2 Test Length

Since an MST is adaptive, it is more efficient than a conventional linear test with respect to test length, meaning that a shortened length of MST often performs as well as a longer linear test while the coverage of test specification is still balanced (Zenisky, 2004). More specifically, previous research findings have shown that even with a reduced test length, MSTs provide equal or increased measurement efficiency and precision compared to linear tests by adapting modules to examinee provisional ability estimates (Hambleton & Xing, 2006; Jodoin et al., 2006; Schnipke & Reese, 1997).

Schnipke and Reese (1997) investigated the use of MST for the Law School Admission Council  (LSAC). They compared the precision of ability estimates obtained from MSTs with a standard CAT design and a linear test (i.e., paper-and-pencil test). The results incidated that all module-based MST designs led to improved precision over the same length linear test and provided almost as much precision as the linear test of double length. Under the context of credentialing exams, Hambleton and Xing (2006) and Jodoin et al. (2006) showed that the performance of an MST design can be effective as much as the linear test. More specifically, Hambleton and Xing (2006) claimed that when a test infomration function (TIF) matches the passing score, the MSTs produced slightly better classification results than the linear test for tests of the same length. Jodoin et al. (2006) observed that a 40-item two-stage test still showed decision accuracy simliar to a 60 item linear test.

In a recent study, Luo and Kim (2018) assembled a test for MST using a top-down and a bottom-up assembly approach with three test lengths (24, 48 and 60) and three panel designs (1-3, 1-2-2, and 1-3-3 MSTs). They found consistent effects of test length on the measurement precision in the simulation study, where longer tests resulted in lower root mean squared errors (RMSEs) regardless of assembly approach and panel designs.

With respect to test length of MST, another consideration for designing an MST is whether content specification should be covered at the module or total test levels (Zenisky, 2004). For example, when content specifications of a test are to be satisfied at the module level, it may require more items at stages and thus, test length increases. Accordingly, test length of an MST should be long enough to produce accurate measurement precision as well as to provide a balanced domain coverage at the module or test levels (Zenisky & Hambleton, 2014).

### 2.1.3 Characteristics of Modules

MST design issues related to characteristics of the modules have been expansively studied because they are closely related to the measurement efficiency and precision of an MST. Generally, two important factors are involved regarding the characteristics of modules: statistical characteristics of modules and the number of items per module.

When a test is assembled using an IRT model, statistical properties of modules is usually characterized by the target module information functions (MIF) to which each module is assembled. The target MIF controls the exact measurement properties of the items selected for module. Therefore, the choice of the statistical target MIF is one of the most important decisions for designing an MST panel, especially when the panel is built using a bottom-up assembly approach (Luecht & Nungester, 1998). To assemble the module-based MST panel, the statistical level of differentiation among the modules within each stage such as the average item difficulty, variance of the item difficulties, and average item discrimination, should be specified (Luecht, 2014). Among all statistical factors, the average item discrimination has a direct effect on the amount of information provided by each module. The average and variance of the item difficulties of each module determine the location and region in the ability scale where each module will cover. Accordingly, a general goal of designing MST is to select the items that approximate the desired level of statistical properties of MIFs, subject to other non-statistical constraints (Hendrickson, 2007).

Each of various studies regarding the statistical characteristics of modules provides informative psychometric results (Jodoin et al, 2006; Kim, Chung, Dodd, & Park, 2012; Kim & Plake, 1993; Zenisky, 2004), but broader conclusions are found in the results. Kim and Plake (1993) argued that the statistical characteristics of the first module for two-stage tests significantly affected the measurement precision. They found that a routing module test with a wide range of item difficulties was superior to the peak routing module test by showing smaller measurement errors, depending on the characteristics of second-stage modules in the test.

Zenisky (2004) also indicated that distributing more information to the first-stage module rather than later-stage modules was recommended for more accurate routing decisions when there was a limited amount of overall test information in the test. This argument was confirmed by Kim et al. (2012). They compared various panel designs of the MST using mixed-format tests in the context of classification testing. In the simulation, the first-stage module was constructed according to three levels of MIFs and three different centers of TIFs. The higher levels of MIFs at the first-stage achieved better accuracy of the classification decision. However, Jodoin et al. (2006) discovered that there were no significant differences of measurement properties between the 1-3-3 MST with the same amount of MIFs across three stages and the 1-3-3 MST with the reduced information for the first-stage and the increased information for the subsequent stages.

To find optimal target MIFs for an MST design, Luecht and Burgin (2003) proposed a way of generating feasible target MIFs, called conditional information targeting (CIT) strategy. This method was used with two purposes: (1) to explicitly control the proportion of the population routed along various pathways and (2) to make the targets as informative as possible, considering the quality of the items in the item bank, content, and other test specifications. Though they illustrated this strategy with a simple 1-2 MST, the simulation results were promising for the CIT strategy to be used for more complicated MST panel designs.

Various number of items per module have been used in MST research and operational testing programs. The module length may vary across the stages, depending on the targeted statistical properties of modules and test specifications of a test (Yan et al., 2014). In fact, the targeted statistical characteristics of modules (i.e., the desired MIFs) and module length are closely, but not directly, related to each other. If the number of items per module is reduced, then it is highly likely to decrease the amount of MIF in the assembled MST unless high-quality items (i.e., items with high values of discrimination) are provided in the item bank.

Both strategies of longer modules at the first-stage and of extended modules at the later stages have their own rationale (Patsula, 1999). The former design is intended to more accurately route examinees to subsequent stages and the later design is needed to provide tailored items to examinees after the test are more closely aligned with the estimates of examinees' abilities. Both strategies may gain some accuracy from one side and lose some accuracy from the other side (Zheng et al., 2012).

Although Kim and Plake (1993) found that increasing the number of items at the first-stage module increased precision, a first-stage module that has too many items may result in a less efficient adaptive test. However, Luo and Kim (2018) found that assigning more items to the final stage resulted in better precision of abiltiy estimation than assigning more items to the first stage when the bottom-up assembly method was used. Even when the panel was assembled with the top-down assembly approach and the controlled routing error, the final stage had many more items and the test produced the best performance of ability estimation among any other MST panel designs.

Interestingly, Zheng et al. (2012) examined allocating a different number of items to module across the stages in three different ways so that the simulated panel designs had equal-length stages with longer earlier stages, longer middle stages, and longer later stages. In the results, it was not clear which allocation strategies had better measurement properties such as classification accuracy and measurement precision.

### 2.1.4 Item Bank and Examinee Population

The design variables described in the previous section play an important role in developing MST panels. But, the extent to which the quality of measurement properties is optimal for any MST design given a specific testing context depends on how well the item bank supports the chosen design of MST and the nature of the examinee population (Zenisky & Hambleton, 2014).

To satisfy the given design variables when building an MST test, the item bank must contain sufficient depth and breadth to facilitate the automated test assembly process. For example, if credentialing test programs expect to move from a conventional linear test format to MST, the item bank that is well suited for a linear test may not be ideal for building the MSTs and, therefore, the theoretical advantages of an MST design would not be realized. This is particularly challenging in MST because the item bank required for the construction of MSTs needs to satisfy the more detailed content and statistical specifications (Hambleton & Xing, 2006; Xing & Hambleton, 2004).

Xing and Hambleton (2004) explored how the size and quality of the item bank affect the psychometric properties of credentialing exams through the comparison of different CAT designs, including a linear test, CAT, and MST. They created two item bank sizes (240 and 480) and manipulated item quality of the bank with three levels (poor, original, and improved levels) by increasing or decreasing the average of item discrimination values. The simulation results showed that the bigger size and better quality of item bank led to significantly better decision consistency and accuracy regardless of test designs.

In a recent study, Wang et al. (2012) further considered two levels of item bank quality which are an operational item bank and an optimal item bank by improving item discriminations and targeting item difficulties in the original bank. The results indicated that the quality of the item bank may be the primary factor that impacts the measurement properties of any MST design. Specifically, They observed that most of MST designs under the optimal items bank were highly effective in terms of achieving appropriate adaption of the module difficulties. Also, the results suggested that although the number of modules per stage increased in the MST panel, the item bank was not able to provide enough items of desired quality and thus, led to less measurement information in specific regions of the scale if the quality of item bank was not optimal.

The examinee population is an important consideration for the measurement properties in MST as well. Though cut-scores for the credentialing or licensure programs can be set independent of the examinee population, it has been found that the characteristics of the population have a clear impact on the process of test assembly and testing results (Zenisky et al., 2010).

As explained in Jodoin et al. (2006), the consistency of correct classification in the pass-fail decision could be affected by the location where more examinees' true abilities are centered. For instance, classification error rates may increase when the passing score is moved from 0.5 to 0.0 in the IRT ability scale if the population follows a standard normal distribution because more examinees are centered near the passing score of 0.0. Also, Hambleton and Xing (2006) investigated whether the TIFs in an MST should be centered to the mean of the candidate ability distribution or to the passing score to maximize the effectiveness of test. When the TIF matched to the passing score, the MST designs resulted in slightly better classification results than the linear test designs.

### 2.2 Evaluation of Test Performance in MST

When assessing test performance, a primary interest of researchers is usually how precisely a test measures the construct of interest. In IRT, the amount of test information, which is simply the sum of the item information for individual items, is directly related to measurement precision of a test (Hambleton, Swaminathan, & Rogers, 1991) because the inverse of square root of the test information given at a value of $\theta$ is the conditional standard error of ability estimate (CSEE) at the $\theta$. In the context of conventional linear test, researchers used the TIF obtained from the test to predict the test performance. Under the adaptive testing context, however, it is challenging to obtain test-level information since the concept of a conventional test form does not apply to CAT and MST (Park et al., 2017). For example, in an MST there exists several different routes that examinees are likely to travel using the testing process. Therefore, simulation methods have been employed in the adaptive testing context to evaluate the performance of a test, which is an appropriate choice of methodology when there is no reasonable analytical way of solving a problem (Harwell, Stone, Hsu, & Kirisci, 1996; Psychometric Society, 1979).

For MST, Monte Carlo (MC) based simulation methods have been a typical methodology used to evaluate the performance of a test for various purposes (e.g., Armstrong et al., 2004; Dallas, 2014; Hambleton & Xing, 2006; Jodoin et al., 2006; Luecht, 2003; Luecht & Burgin, 2003; Luo & Kim, 2018; Park et al., 2014; Park et al., 2017; Wang, 2017; Wang et al., 2012; Weissman et al., 2007; Xing & Hambleton, 2004; Zenisky, 2004; Zheng et al., 2012). For example, some simulation studies have focused on comparing psychometric properties of an MST with other testing designs such as a linear test and a CAT based on the operational test setting (e.g., Hambleton & Xing, 2006; Jodoin et al., 2006; Wang, 2017). Some studies have investigated impacts of varying design variables described in the previous section on the measurement precision of MST (e.g., Wang et al., 2012; Zheng et al., 2012). Other studies have conducted a simulation study to examine the performance of new method related to the implementation of MST (e.g., Luo & Kim, 2018; Park et al., 2014; Park et al., 2017; Weissman et al., 2007). Generally, the MC based simulation methods use the following procedures to assess measurement precision under an MST context. First, item parameters are drawn either from a real item bank or from underlying item parameter distributions. Second, ability values are generated from a specific population distribution (e.g., a standard normal distribution). Third, simulate examinees' responses for each route in a specific MST design following an underlying IRT model. Fourth, examinee abilities are estimated using a predetermined scoring method (e.g., MLE). Fifth, the whole process is replicated $R$ times, and the ability estimates obtained from $R$ replications are used to calculate, for instance, the CSEEs along the $\theta$scale.

Recently, Park et al. (2017) proposed an analytical approach to assess performance of an MST. Not only did they derive a method to compute the CSEEs but they also suggested a way of predicting classification accuracy for MST using the CSEEs. To compute the standard error given a specific $\theta$, test-level information for MST (i.e., MST test information) is derived using two steps: (1) construction of test information for primary routes and (2) averaging of the test information to take secondary routes into consideration. In the first step, a single TIF is formed only using the peak of each other primary pathway information functions along the $\theta$ scale. In the second step, the single TIF is averaged across the $\theta$ scale using a simple moving average method to incorporate the secondary pathways into the calculation of the MST test information. Then, the averaged MST TIF is used to obtain the CSEEs. Park et al. showed that the analytically computed CSEEs and the predicted classification accuracy were close to those obtained from a simulation.

In the study, Park et al. (2017) compared the performance of the proposed analytical and the simulation-based evaluation methods in the MST context. They used a real item pool from the science subset of the 1996 National Assessment of Educational Progress (NAEP). Examinee abilities were estimated using the expected a priori (EAP) scoring method with a standard normal prior. To compare the measurement precision of both analytical and simulation-based approaches, four design variables were manipulated:

two levels of test length (40 and 60 items), two MST configurations (1-2-2- and 1-3-3 MSTs), two routing module designs, and four proportions between dichotomous and polytomous items. To predict the classification accuracy, three cutoff scores were applied. In the results, both the suggested analytical and simulation evaluation methods produced similar standard errors along with ability scale and classification accuracy to each other, leading to conclusion that the analytical evaluation method based on the MST test information effectively predicted the performance of MST.

Park et al. (2017)'s study is meaningful for two reasons. First, the analytical approach can help researchers assess the performance of MST more efficiently. It usually takes significantly more time and effort to set up and conduct a simulation study, especially as the number of conditions and replications become large. However, since the analytical evaluation method does not require multiple replications, the computing process is relatively fast. Second, the evaluation results of analytical method may be more exact than those of the simulation method. For example, since the Park et al. (2017) study sampled ability parameters from a normal distribution for the simulation-based approach, they observed large standard errors for extreme ability parameters that was mainly due to the small number of simulees. However, the analytical method is able to calculate the measurement precision without relying on the number of examinees at the extreme ability level.

Although the proposed method by Park et al. (2017) can be analytically derived, a part of its derivation is open to question. First, the averaged MST test information may not be accurate because the simple moving average method does not use the exact test information values from secondary routes. As the MST panel contains more secondary routes (e.g., 1-3-5 MST), the performance of the analytical method in Park et al. (2017) would be more doubtful. Second, they computed the simple moving average of 5 points around each $\theta$ value without a specific reason for using the 5 points. From a small simulation study, however, it was found that the MST test information varied depending on the number of the points being used to calculate the simple moving average. Third, it is questionable whether applying the simple moving average method is appropriate for two-stage MSTs such as 1-2 and 1-3 MSTs because no secondary route exists in the two-stage MSTs. Obtaining the MST test information only from the primary routes does not make sense as well because a routing is not perfectly accurate due to measurement error of ability estimation in practice.

In this study, a new analytical evaluation approach, which is based on the equated NC scoring method (Stocking, 1996), is proposed to calculate the CSEEs. This method can overcome the drawback of Park et al. (2017)'s approach because measurement precision is more exactly derived. The proposed method is fully discussed in Chapter 3.

## 2.3 Test Assembly

In MST, a test is built using either a top-down assembly method or a bottom-up assembly method (Luecht & Nungester, 1998). Once the assembly strategy is decided, the test is usually automatically assembled using either a linear programming approach or a heuristic approach. In this section, the top-down and bottom-up assembly strategies are introduced first. Then, the automated test assembly (ATA) methods are discussed.

### 2.3.1 Top-Down and Bottom-Up Approaches

In MST, there are two general test assembly strategies to build panels: the bottom-up and the top-down approaches (Luecht & Nungester, 1998). The bottom-up approach is considered a "divide-and-conquer" method because a test level specification for an MST design is decomposed into the module level specification (Luo & Kim, 2018). Thus, it is necessary to prepare comprehensive module level specifications for the statistical and non-statistical targets so that each module can be assembled independently across stages. To develop multiple parallel panels, multiple parallel modules are assembled first and then they are mixed and matched because modules representing the same difficulty level are exchangeable across panels.

In contrast, the top-down strategy is a holistic test assembly approach in that test level specifications for the statistical and non-statistical targets are constrained on routes (Luo & Kim, 2018). Therefore, test developers do not need to make specific specifications for modules. Instead, the top-down approach lets the computer algorithm determine the best partition of the test level constraints across stages. Accordingly, modules can be built in prescribed ways at each route to satisfy the desired test level properties (Luecht & Nungester, 1998). Note that, since the test level specifications are imposed, modules representing the same difficulty level may not be completely parallel and exchangeable across panels when the top-down strategy is used. Luecht and Nungester (1998) stated that employing the primary routes in the panel of MST to develop target TIFs is enough because the secondary routes would not likely to be used for most examinees, whereas Zheng et al. (2012) and Luo and Kim (2018) used the target TIFs of all routes in the ATA process.

Until now, most MST studies have employed the bottom-up strategy (e.g., Hambleton & Xing, 2006; Jodoin et al., 2006; Luecht & Burgin, 2003; Luecht, Brumfield, & Breithaupt, 2006; Park et al., 2017; Wang, 2017; Weissman et al., 2007; Zenisky, 2004; Zenisky et al., 2018) because the process of the multiple module assembly is similar to that of building multiple linear test forms which might be a more familiar test design concept to test developers (Luo & Kim, 2018) and it is more straightforward to implement compared to the top-down strategy (Dallas, 2014; Zheng, Wang, Culbertson, & Chang, 2014). However, when test developers try to find an optimal MST design given a specific testing context, the task is very difficult under the bottom-up approach because there are too many design variables to be considered at the module level. For example, we need to decide the number of items within each module at each stage and how to distribute test information across stages to maximize test score precision. As the number of modules and stages increase in the MST configuration, the number of design variables that should be taken into account to decompose the test level specification into the module level increases exponentially. Because there is no analytical method to find the optimal partition of test level design variables that guarantees the best measurement precision under the bottom-up strategy, the distribution of design variables into the module level usually depends on experiment or the test developer's experience (Luo & Kim, 2018). Therefore, the design space, where test developers can search to find the optimal design of MST, is limited in practice.

However, the top-down approach does not require artificial decisions to find an optimal decomposition of test level design variables into the module level because the ATA algorithm can find the best solution for the partition by meeting all constraints as well as ensuring the best measurement precision. Due to the simplification of the designing process and the flexibility to set constraints across stages under the top-down assembly, the test specification of the conventional linear test form can be easily moved into an MST and better psychometric characteristics (e.g., measurement precision) can be obtained compared to the bottom-up approach (Luo & Kim, 2018).

Compared to the bottom-up approach, there has been only few studies on the top-down assembly strategy in MST (Luo & Kim, 2018; Zheng et al., 2012). Zheng et al. (2012) investigated the feasibility of the top-down assembly in MST for a larage-scale classification test. They assembled each panel by using a real item bank with 600 items and a revised version of the normalirzed weighted absolute deviation heuristic (NWADH; Luecht, 1998) method. In the ATA process, they used two steps: (1) assembled multiple modules from the item bank without a full constraint on content category and then (2) built MST panels from the obtained modules. Because the content specifications were constrained at the test level, they applied the top-down strategy at the second step to monitor the quality of the assembled panels. However, the assembly approach used in Zheng et al. (2012) should be considered a hybrid assembly method rather than a full top-down approach. This is because only content constraints were set at the test level wheresas other statistical and nonstatistical specifications (e.g., the number of items in each module and target MIFs) were imposed at the module level. In a simulation study, they compared the performance of an MST under the top-down approach with that of a linear test and a CAT. The results showed that the MST had better classification accuracy than the linear test and provided classification accuracy as good as that from CAT.

Recently, Luo and Kim (2018) proposed a route -based top-down assembly strategy using mixed integer linear programming (MILP; van der Linden & Adema, 1998). In this approach, the design variables on all allowed routes were constrained at the test level. To control routing errors, they divided the population into subpopulations based on the routing decision points (RDPs) and then set two types of constraints at the module level: (1) to anchor the RDP between two adjacent modules and (2) to set the minimum information at the RDP for modules. In an ATA process under the bottom-up approach, the objective is usually to make the observed MIF as close as possible to a given target MIF.  Under the top-down approach of Luo and Kim, the objective was to maximize all possible routing information functions (RIF) over the $\theta$ interval of targeted subpopulation. Through the simulation study, they compared the top-down and the bottom-up MST designs in terms of measurement precision and the route usage rate using three panel configurations (1-3, 1-2-2, and 1-3-3 MSTs). The results indicated the top-down approach had higher measurement precision and better controlled routing error than the bottom-up approach.

### 2.3.2 Automated Test Assembly (ATA)

Generally, two ATA approaches are widely used in MST: heuristic and mixed-integer programming (MIP). A heuristic-based ATA approach builds a test by solving a series of local optimization problems to select a best-fitting item or set of items in the process of test assembly (Zenisky, 2004; Zheng et al., 2014). At each sequence of the optimization problem, a composite objective function is used to meet the statistical targets (e.g. TIF) and non-statistical constraints (e.g., test specification for content).

Various heuristic test assembly methods have been developed such as the weighted deviation method (WDM; Swanson & Stocking, 1993), the maximum priority index (MPI; Cheng & Chang, 2009), and the normalized weighted absolute deviation heuristic (NWADH; Luecht, 1998) in the adaptive testing context. Among them, only the NWADH has been adapted in many MST studies (e.g., Hambleton & Xing, 2006; Jodoin et al., 2006; Luecht & Burgin, 2003; Luecht & Nungester, 1998; Patsula, 1999; Zheng & Chang, 2015; Zheng et al, 2012). This heuristic method normalizes the weighted deviations for constraints to put them on a common scale. The item with the smallest normalized absolute deviation is selected for a test (Zheng et al., 2014). The heuristic-based ATA methods provide a reasonable solution for the test assembly in relatively short period of time, but do not guarantee that the assembled test forms exactly satisfy all of the constraints in the test specification and the obtained solution may not be the best possible solution (Luecht, 1998; Breithaupt & Hare, 2007).

A second popular ATA approach is to use the MIP algorithm (van der Linden, 2000, 2005) in MST. In the MIP approach, all the ATA problems (e.g., constraints and objective functions) are translated into a set of mathematical linear formulae and a solution for assembling test forms/panels can be found which optimizes the objective functions. The solution in the MIP is the best possible solution among a large number of feasible solutions. However, as the complexity and the number of constraints increase (e.g., increase of the number of parallel test forms), the ATA procedure can be time-consuming. In addition, when the current item pool does not suffice to meet all of the constraints, then the algorithm may return no solution because the ATA problem is infeasible in that case (van der Linden, 2005).

The ATA problem using MIP for any assembly of a test can be solved by taking four steps introduced by van der Linden (2005): (1) identify the decision variables; (2) model the constraints; (3) model the objective; and (4) solve the model for an optimal solution. In the first step, the decision variables consist of a 0 or1 (0/1) and a few real-valued (continuous) variables are defined for the MIP. The 0/1 variables are used to identify whether each item is selected in the assembled test form. The real-valued variables are necessary for technical reasons such as defining an objective function using test information. A set of constraints and objectives on the selection of a test form is formulated based on the defined decision variables (Breithaupt & Hare, 2007).

In the second step, all of the constraints are expressed as a form of inequality/equality that imposes a real-valued bound on a (un)weighted sum of decision variables. There are three types of constraints that depend on their attributes: quantitative, categorical, and logical constraints (van der Linden, 2005). The quantitative constraints set a real-valued bound on a weighted sum of the decision variable and are related to the quantitative attributes of item/test such as item/test information, test length, and expected response times. The categorical constraints impose an integer bound on an unweighted sum of decision variables and usually deal with content category, item format, and cognitive level of item. To formulate the categorical constraints, the item pool is partitioned into subsets of items with the same attributes. The logical constraints are to control the logical relation between pairs, triples, and so on, of items. For example, this constraint prevents the selection of one item if the other is selected from a set of “enemy” items.

In the third step, an objective function can be defined using any attribute of constraints formulated as a mathematical expression in step 2 (van der Linden, 2005). If the IRT model is used in the ATA process, a general form of the objection function is to maximize the TIF or minimize the deviation of the test information of the assembled test from the target information (Zheng et al., 2014). Note that a test assembly problem can have more than one objective function if a set of fixed ability values are used. Once all of the ATA problems are set up through step 1-3, the MIP formulation is transferred into software programs, known as solvers, to find an optimal solution. Currently, many different solvers are available, including open source solvers (e.g., lp_solve; Diao & van der Linden, 2011) and commercial solvers (e.g., IBM ILOG OPL; Luecht et al., 2006). The decision variables recorded as 1 in the solution indicate the chosen items for the assembled test form.

The following is a standard model formulation to assemble a single test form with a quantitative objective from a discrete item pool. The model involves maximizing test information at several fixed ability values subject to a test length, content constraints, and enemy item specifications. The items in the pool are denoted as $i = 1, \dots, I$. Let $V_c$ be a subset of items in the pool that belong to content category $c$ and $n_c$ is the lower bound on the number of items from this subset. $V_e$ is a subset of items that belong to an enemy item set $e$. Also, $I_i(\theta_k)$ indicates the information function value for item *i* at *k*th $\theta$. Now decision variables for items in the pool are defined as

$$x_i = \begin{cases} 1 & \text{if item } i \text{ is selected for a test form,} \\ 0 & \text{otherwise} \end{cases} \tag{2.3.1}$$

The model has an objective function based on the maximum principle, which maximizes a minimum value of the TIFs among the fixed $K$ values of $\theta$, such that:

$$\max\ y \text{ (objective function)} \tag{2.3.2}$$

subject to possible constraints as followings:

$$\sum\nolimits_{i=1}^{I} I_i(\theta_k)x_i \geq y, \quad \text{for all } k, \tag{2.3.3}$$

$$\sum\nolimits_{i=1}^{I} x_i \ \text{¦}\ n \text{ (total test length)}, \tag{2.3.4}$$

$$\sum\nolimits_{i \in V_c} x_i \ \text{¦}\ n_c \text{ (content category)}, \quad \text{for all } c, \tag{2.3.5}$$

$$\sum\nolimits_{i \in V_e} x_i \leq n_e \text{ (mutually exclusive items)}, \quad \text{for all } e, \tag{2.3.6}$$

$$x_i \in \{0, 1\}, \quad \text{for all } i, \tag{2.3.7}$$

where $y$ is a real-valued decision variable in Equation (2.3.2). A symbol of $\gtreqless$ indicates the choice of an equality or an inequality sign. Equation (2.3.5) is a constraint for the content category requirement in a test form. Equation (2.3.6) guarantees no item overlap among a set of enemy items. A fixed test length is imposed by Equation (2.3.4). This standard ATA model formulae can be flexibly reformulated depending on testing purposes, test designs, assembly strategies, specific constraints, and other factors.

Since Adema (1990) proposed MIP models for constructing two-stage MST, the MIP method has become one of the popular strategies for test assembly in MST. van der Linden (2005) dealt with comprehensive details about linear modeling of ATA process for various test designs, including the MST and Diao and van der Linden (2011) described how to use the free solver of lp_Solve with R interface to conduct the ATA for MST. Some studies discussed the assembly of MST using MIP technique in an operational context such as the Law school Admission Test (LSAT) (Armstrong et al., 2004; Armstrong & Roussos, 2005; Weissman et al., 2007) and the Uniform Certified Public Accountant (CPA) examination (Breithaupt, Ariel, & Hare, 2010; Breithaupt, Ariel, & Veldkamp, 2005; Breithaupt & Hare, 2007; Luecht et al., 2006; Melican et al., 2010). Recently, Park et al. (2014) adoped the MIP method for constructing an MST using the mixed-format test to enhance item pool utilization and Luo and Kim (2018) applied the MIP model to suggest a top-down assembly in MST.

Under MSTs, MIP algorithm can provide an optimal assembly of parallel panels that strictly satisfy all constraints at the module level or test level. Because it is required to assemble multiple panels and modules for practical reasons (e.g., test security) in MST, however, this makes the ATA problem in MIP more complicated. Thus, one may need to use a high-performing solver in the ATA process or tune MIP parameters to increase efficiency of problem solving (Luo & Kim, 2018).

### 2.4 Scoring and Routing Methods

Among many critical considerations in the development of an MST, the choice of strategies for scoring and routing has enormous implications for testing results. In this section, IRT pattern-based scoring and IRT summed score-based scoring methods are discussed. Following the scoring methods, two routing methods of the approximate maximum information (AMI) and the defined population interval (DPI) are reviewed next.

### 2.4.1 IRT Pattern-Based and IRT Summed Score-Based Scoring Methods

In CAT, IRT pattern-based scoring methods such as the maximum likelihood (ML) estimation and Bayesian expected a posteriori (EAP) estimation are typically used to estimate examinee ability $\theta$. The IRT pattern-based scoring methods use all information available in an examinee's response pattern under IRT models and therefore, each response pattern is often associated with a unique ability estimate $\hat{\theta}$ (Thissen, Pommerich, Billeaud, & Willams, 1995). Despite this advantage, the pattern-based scoring requires a complicated process when estimating examinee ability. For example, the ML estimate of $\theta$ is obtained by finding a $\theta$ value which maximizes the likelihood of a probability function that an examinee's response pattern for a set of items would be observed. Due to this property of the ML estimation, examinees who respond to the same set of items and have the same summed score, but had a different responses pattern will receive differeent ability estimates. This process of scoring is not intuitive to test takers who are familiar with a scoring method where a test score is a sum of the number of items answered correctly. The scoring feature of ML estimation makes it even more challenging to explain test takers a system of deriving the test scores in the adaptive testing (Stocking, 1996).

Those responsible for testing programs should provide score interpretations appropriate to the audience in simple language when releasing test score information (American Educational Research Association, American Psychological Association, and National Council for Measurement in Education, 2014, p. 119). For this purpose, the use of a scoring method based on the summed score (or NC score) of a test is more desirable than the response patterns-based scoring methods in terms of score interpretation. In practice, some conventional standardized testing where linear test forms are administered have been implemented with the IRT scaled summed-score methods (Stocking, 1996; Thissen et al., 1995). Two popular IRT scaled summed scoring methods under the conventional linear testing design are: the equated NC (ENC) scoring, which is based on the inverse of test characteristic curve (TCC) (Kolen & Brennan, 2004; Stocking, 1996), and the EAP summed scoring, which is based on the compound binomial distribution of item response probabilities for each summed score (Thissen & Orlando, 2001; Yen, 1984). Under MST, the ENC scoring method has been frequently used to determine ability estimates when routing examinees to the next module (e.g., Luecht et al., 2006) but, only few studies have employed the IRT scaled summed-scoring method for the final estimation of examinee abilities (e.g., Kim & Moses, 2014). In practice, some of the current comprehensive testing programs apply the ENC scoring as a final scoring method in the large-scale statewide assessments under the MST context (e.g., Wendler & Bridgeman, 2014).

To use the IRT scaled summed score methods in adaptive testing, including MST, the following three conditions must be met (Stocking, 1996):

(1) Whether the interpretability of the scoring is enhanced?

(2) Whether the scoring is sensitive to test difficulty of different test forms?

(3) Whether the scoring can accomplish both (1) and (2) with accurate measurement precision.

As already discussed above, the first condition is easily satisfied because a sum of number correct answers in a test corresponds to each IRT scaled summed-score and all items count the same amount towards this score. Accordingly, this scoring system is more understandable to test takers. The second condtion is related to placing test scores from different parallel test forms onto the same score scale, called equating, as in the conventional linear tests (Kolen & Brennan, 2004). Because MST is adaptive, which tailors the difficulty of next module to examinee’s current ability estimate, test takers with different proficiencies may have different routes at the end of testing. Therefore, raw summed scores from different routes would not be comparable since the different pathways could differ by difficulty. However, the IRT scaled summed-scores are estimated by using item parameter estimates from the item pool where all item parameters are already calibrated on the same scale. Thus, the produced score is eventually a result of a scaling that adjusts for form-to-form variation in test difficulty (Stocking, 1996).

To satisfy the last condition, the IRT scaled summed-score should be comparable to the IRT pattern-based estimates in the sense of measurement accuracy and efficiency. Fortunately, several studies have shown that the summed score based IRT scaled estimates can be used to obtain a good estimate of examinee’s proficiency (Chen & Thissen, 1999; Kolen & Tong, 2010; Thissen et al., 1995; Thissen & Orlando, 2001; Stocking, 1996; Yen, 1984).

Yen (1984), Thissen et al. (1995), and Thissen and Orlando (2001) proposed a way of estimating the latent ability corresponding to each summed score by demonstrating the relationship between the summed score and response pattern score. Yen (1984) used an approximation method to calculate the compound binomial distribution of item response probabilities given a summed score (Lord & Novick, 1968) and found the NC maximum likelihood ability estimate corresponding to that summed score. In her study, the ability estimates provided accurate estimated true scores for a group associated with each summed score. Instead of the approximate ML estimation, Thissen et al. (1995) and Thissen and Orlando (2001) used the EAP summed score estimation method which is the mean of a posterior density for a summed score, given the item parameter estimates. Both studies noted that effectively computed IRT scaled scores for each summed score is useful for score reporting, though a small loss of information is inevitable due to the simplification of scoring from response patterns to summed scores. Therefore, the IRT scaled score associated with the summed score can provide a reasonable latent ability estimate (Thissen & Orlando, 2001).

Stocking (1996) and Kolen and Tong (2010) also demonstrated that the IRT scaled summed score methods are comparable with the IRT pattern scoring methods in terms of measurement accuracy. Stocking (1996) explored whether the ENC scoring method could be employed without undue sacrifices to the other efficiencies gained from the IRT pattern-based scoring method (e.g., MLE) under the adaptive testing context. In this scoring method, the IRT scaled ability estimate is obtained by finding an inverse value of the NC score on the IRT ability scale through the TCC. Stocking showed from a simulation study that the ENC scoring could be a feasible alternative to the ML estimation although it has some reduced information compared to the full information scoring approach in CAT. Specifically, measurement precision of the ENC scores in CAT was judged acceptable since both the ENC scoring and ML estimation methods produced very similar reliability and CSEE curves along the$\theta$scale.

Kolen and Tong (2010) compared four IRT pattern-based scoring methods, MLE and EAP, and two IRT scaled summed scoring methods, the ENC scoring, and EAP summed scoring methods. The study showed that the score distributions between the IRT scaled summed score (the ENC scoring and EAP summed scoring) versus pattern-based score (MLE and EAP) estimators were very similar. This indicates that it is difficult to argue whether pattern-based estimators (e.g., MLE) are superior to summed scoring methods (e.g., the ENC scoring) in a particular application (Kolen & Tong, 2010). Therefore, they argued that the use of ability estimation methods based on summed scores might be appropriate for those testing programs where it is important to have a simple explanation for scoring.

In addition, Chen and Thissen (1999) estimated item parameters of the 3PL model by modifying the maximum marginal likelihood (MML) EM algorithm with the computations based on summed scores instead of response patterns. Although Chen and Thissen's scoring requires a complicated procedure compared to other IRT scaled summed score methods, the result showed that the estimated scaled scores were approximately as accurate as those obtained using the pattern-based ML estimates.

As all previous research noted, it seems that IRT scaled summed score-based method can be an effective alternative scoring in MST context and provides (1) ease of score interpretation, (2) comparability of scores from different routes, (3) acceptable measurement efficiency and precision. As other studies indicated, a loss of information is inevitable due to the simplification of scoring for the IRT scaled summed score-based scoring (Stocking, 1996; Thissen et al., 1995; Thissen & Orlando, 2001). But, the amount of loss is small, and more importantly, the loss may be counterbalanced by a practical advantage of the summed score-based scoring strategies, obviously easiness of interpretability of scores when the scoring report is released to public (Thissen et al., 1995). Despite of some loss of information, the summed score-based scorings also produce measurement estimation that are more robust to suboptimal test adaptation under the adaptive testing context or misleading responses due to other factors, such as misunderstanding the directions, anxiety, and poor time management (Meijer & Nering, 1997; Stoking, 1996; Stocking, Steffen, & Eignor, 2002).

### 2.4.2 AMI and DPI Routing Methods

In addition to the scoring method, another important decision that must be made in MST is what method to use for routing. The routing is a process that assigns an examinee to a well-matched module at the next stage based on the examinee's performance on the previously selected module(s). A choice of routing rule usually depends on the testing purpose and design of MST (Yan et al., 2014) and it affects the usefulness of the results from the MST (Zenisky & Hambleton, 2014).

Two routing rules are commonly used for MST: the approximate maximum information (AMI; e.g., Luecht et al., 2006) method and the defined population interval (DPI; e.g., Jodoin et al., 2006) method. The AMI method uses a point where two empirical adjacent MIFs at the next stage intersect as the routing decision point (RDP). Once the RDP is found, examinees are assigned to the module which has maximum information at their current ability estimate. This method is similar to using a maximum information criterion to select an item in CAT, given a current provisional estimate. The intersection point can be found using a numerical root-finding method such as the bisection method. Since the AMI is the IRT information-based method, the module with maximum information could be found using either the IRT pattern-based or IRT scaled summed score-based scoring methods. For IRT scaled summed score-based scorings, typically the ENC scoring method has been used and this method has performed as well as the IRT pattern-based scoring method (Dallas, 2014; Luecht et al., 2006; Zheng et al., 2012).

The DPI rule uses the prespecified proportion of examinees in the population distribution to find the RDP. For example, if we want the approximately equal module usage at the second stage in a 1-3 MST panel, the two RDPs can be found at two ability points associated with 33rd and 67th percentiles in the cumulative population distribution. This method is usually used to manage the module usage rate (Hambleton & Xing, 2006; Luo & Kim, 2018). Although those two routing methods have different purposes, Dallas (2014) found that the AMI approach performed better than the DPI approach in terms of measurement precision.

# CHAPTER 3

# STUDY 1

The dissertation consists of two studies. Study 1 introduces a new method to analytically evaluate MST performance without conducting a simulation and demonstrates that the proposed method provides accurate estimates of measurement precision and classification accuracy in an MST. In Chapter 4, Study 2 proposes a process of finding an MST design that has optimal measurement properties measured by the analytical evaluation method, given a specific set of test circumstances.

## 3.1 Analytical Evaluation of MSTs

A concise and effective index of a test's measurement properties in estimation of examinee ability is the CSEE. The CSEEs provide evidence of how accurately a test measures the examinees' proficiencies across the ability scale (Hambleton et al, 1991). In IRT, the CSEE is obtained by transforming the test information given an ability $\theta$ using a formula as such,

$$CSEE(\theta) = \frac{1}{\sqrt{I(\theta)}}, \tag{3.1.1}$$

where $I(\theta)$ is the test information at $\theta$. As stated in the previous chapter, it is challenging to analytically compute the test information under the adaptive testing because the adaptive tests usually have non-parallel multiple test forms and examinees administer different forms of the test depending on their proficiencies.

Park et al. (2017) developed an analytical method to compute the MST test information using the IRT pattern-based scoring method (e.g., MLE). Based on the MST test information, they showed that the computed CSEEs and the predicted classification accuracy were close to those obtained from a simulation. However, the derived MST test information by Park et al. (2017) may not be exactly accurate for several reasons as mentioned in Chapter 2. First, the exact test information values of the secondary routes are not used in the calculation process of the MST TIF. Second, it is uncertain why the five points around each $\theta$ are used when computing the simple moving average. Third, when there is no secondary route for the two-stage MSTs, applying the simple moving average method is questionable.

In this study, a new analytical evaluation approach, which is based on the ENC scoring method (Kolen & Tong, 2010; Stocking, 1996), is proposed to derive measurement precision (i.e., conditional bias and CSEE). This method can overcome the drawback of Park et al.'s (2017) approach because measurement precision is more exactly derived, implying that when an infinite number of examinees with the same proficiency is used in a simulation study, the estimated CSEE from the simulation will converge to the analytically obtained CSEE. Accordingly, the evaluation results of MST from the proposed analytical method may be more exact than those from the MC-based simulation method since the analytical method is able to compute measurement precision without relying on the number of examinees at each $\theta$.

In addition to measurement precision, many licensure or certification testing programs are interested in how to accurately classify examinees into categories according to their performance levels. Therefore, classification accuracy is an important element of measurement properties when MST is applied to credential testing programs. As indicated in Park et al. (2017), the classification accuracy of an MST can be predicted analytically by employing the method developed by Rudner (2001, 2005) using the

CSEEs at a grid of discrete $\theta$s. Therefore, this study analytically derives the classification accuracy of an MST using the CSEEs obtained from the new analytical method as well.

This study conducts two simulation studies to demonstrate that the new analytical evaluation method is able to predict MST performance more accurately than the MC-based simulation method. Specifically, the measurement precision and the classification accuracy from the simulation method with the ENC scoring were compared with those from the proposed analytical method. In addition, the CSEEs and the classification accuracy obtained from Park et al.'s (2017) analytical approach were also compared with those from the simulation method with the MLE in the two simulation studies to show that the new analytical method is more exact. Therefore, the MST performance is assessed using four different methods in this study: the new analytical method, Park et al.'s (2017) method based on the MST TIF, two MC-based simulation methods which are based on the ENC and MLE scorings, respectively.

Study 1 is organized as follows. First, Park et al.'s (2017) MST test information approach to computing the CSEE is reviewed. Second, the new analytical evaluation method for MSTs is explained in detail. Specifically, the ENC scoring method is introduced first and then, the analytical approach to computing measurement precision of an MST is described. Third, the analytically derived CSEEs are then applied to predict the classification accuracy for an MST. Fourth, two simulation studies were carried out to demonstrate that the proposed evaluation method based on the ENC scoring performs well in assessing measurement properties of an MST. Last, the advantages and implications of the proposed method are discussed.

### 3.1.1 Park et al.'s (2017) Analytical Approach

Park et al. (2017) suggested an analytical method to derive an MST test information function based on the IRT pattern-based scoring methods. Two steps are necessary to compute the MST test information. Suppose that the MST test information for the 1-3-3 MST (see the right panel in Figure 1) is derived. In the first step, a single TIF $I_{prime}(\theta)$ is obtained from the top of three primary route information functions (RIFs) on the $\theta$ scale as follows:

$$I_{prim}(\theta) = 1(\theta < C_1)I_{easy}(\theta) + 1(C_1 \le \theta \le C_2)I_{med}(\theta) + 1(\theta > C_2)I_{hard}(\theta), \qquad (3.1.2)$$

where $1(x)$ is an indicator function equal to 1 when the condition $x$ is satisfied and 0 otherwise; $I_{easy}(x)$, $I_{med}(x)$, and $I_{hard}(x)$ are the three primary RIFs; and $C_1$ and $C_2$ are the points where two adjacent primary RIFs intersect.

In the second step, a simple moving average method is applied at each ability point on a grid of the $\theta$ scale to take the secondary routes into consideration in the derivation of the MST test information $I_{MST}(\theta)$. Park et al. (2017) used the simple moving average of 5 points expressed as:

$$I_{MST}(\theta) = \frac{\sum_{i=1}^{5} I_{prim}(\theta + i - 3)}{5}. \qquad (3.1.3)$$

Then, the inverse of square root of the $I_{MST}(\theta)$ is the CSEE at a specific $\theta$ value.

### 3.1.2 New Analytical Approach

A key to the new analytical evaluation method is the use of ENC scoring (Kolen & Brennan, 2004; Lord, 1980; Stocking, 1996) for ability estimation. Based on the ENC scoring method, measurement precision of MST can be analytically derived using a recursive algorithm (Lord & Wingersky, 1984). This section describes how to compute the conditional bias and CSEE at a specific $\theta$ value in detail.

### 3.1.2.1 Equated Number-Correct (ENC) Scoring Method

The ENC scoring method provides similar results of ability estimation to more common IRT pattern-based scoring of MLE (Kolen & Tong, 2010; Stocking, 1996) and allows performance of an MST to be evaluated without recourse to data simulation. The ENC scoring method estimates an examinee's ability $\theta$ by mapping a NC score (or observed score) "backward" through the TCC using Equation (3.1.4) defined as,

$$\xi(\theta) = \sum_{i=1}^{n} P_i(U = u \mid \theta; \gamma), \tag{3.1.4}$$

where $\xi(\theta)$ is referred to the NC score in a test for an examinee with the $\theta$, $\gamma$ is a vector of item parameters, and *n* is test length ($i = 1, 2, \ldots, n$). $P_i(K = k|\theta;\gamma)$ represents the IRT category characteristic function which indicates the probability of earning a category score *k* on item *i* with the $\theta$. To find a value of $\theta$ in the IRT ability scale corresponding to $\xi(\theta)$, a numerical iterative process such Newton-Raphson or bisection methods is usually employed. Note that to avoid the ability estimates of $\theta$s for a zero and perfect NC scores have infinite values, a small value (e.g., 0.5) is added and subtracted from the zero and perfect NC scores. Also, the range of IRT ability scale is restricted in $[-0.5, 0.5]$ so that the ability estimates do not have very extreme values.

In addition, when employing the IRT three-parameter logistic (3PL) model in a test, ability estimates of $\theta$s for NC scores less than the sum of item guessing parameters *c* are not attainable. This is because the probability of correct answer on an item of the 3PL model asymptotically approaches the value of $c$ as $\theta$ approaches $-\infty$. In this case, a range of NC scores ($\xi$s) where the corresponding $\theta$ can be estimated is given by:

$$\sum_{i}^{n} c_i < \xi \le n. \tag{3.1.5}$$

To find the ability estimates for the NC scores outside the range in Equation (3.1.5) when the IRT 3PL model is used, an ad hoc procedure is needed. In this study, a linear interpolation method is used with the restricted range of possible $\theta$ values. The ad hoc procedure is as follows:

(1) Restrict a range of the IRT $\theta$ scale. In this study, $-5.0$ and 5.0 are used for the minimum and maximum $\theta$s in the range, respectively. Let those two $\theta$s be $\theta_{min}$ and $\theta_{max}$. Any estimated $\theta$s for the NC scores less than $\theta_{min}$ are forced to $\theta_{min}$. Similarly, any estimated $\theta$s beyond $\theta_{max}$ are constrained to $\theta_{max}$.

(2) If the IRT 3PL model is used, find a NC score greater than the sum of the item guessing parameters $c$ in the test form. For example, if the sum of the guessing parameters is 2.3, the NC score 3 is a possible minimum NC score whose corresponding $\theta$ can be found through the inverse of the TCC. Let the NC score be $X$ and the corresponding $\theta$ be $\theta_X$.

(3) Use linear interpolation to find a value of $\theta$ for NC scores between $\theta_{min}$ and $\theta_X$.

To formulize this procedure, define $\theta^*$ as an ability estimate corresponding to a NC score $Y$ between a zero and $X$. The value of $\theta^*$ then are defined by the following equations:

$$\theta^* = \frac{Y-\beta}{\alpha}, \quad where \;\; 0 \le Y \le X,$$
$$\alpha = \frac{X}{\theta_X - \theta_{min}}, \;\; and \;\; \beta = -\alpha\theta_{min}. \tag{3.1.6}$$

Note that step 2 and 3 are necessary only when the IRT 3PL model is employed, otherwise only step 1 is implemented.

#### 3.1.2.2 Derivation of Measurement Precision of MSTs

In IRT, a common measure of precision for ability estimation is the CSEE, which is the standard deviation of the distribution of estimated ability around a true ability. The CSEE is therefore computed uniquely at each true ability value. Under an adaptive testing, a popular way of obtaining the CSEE is to simulate thousands of tests, all based on the same true ability value. This produces a distribution of thousands of ability estimates, roughly centered on the true ability value. The CSEE for the true ability is then the standard deviation of the distribution of estimates.

Although MST performance could be evaluated by estimating the CSEEs through the simulation, many performance characteristics of the MST can be assessed by what is essentially an analytic method. The key to this method is the recursive algorithm suggested by Lord and Wingersky (1984). They developed this algorithm to generate the distribution of observed NC scores for examinees of a given ability $\theta$ using IRT models. Suppose that a test consists of $n$ dichotomous items and the assumption of local independence of IRT is satisfied. Denote the probabilities of answering correctly on the $n$ items in the test as $P_1(\theta_t), P_2(\theta_t), \ldots, P_n(\theta_t)$, where $\theta_t$ is a particular ability value. The probabilities of two possible NC scores (0 and 1) for the first item are given by

$Q_1(\theta_t) = 1 - P_1(\theta_t)$ and $P_1(\theta_t)$. Adding the second item to the first item now allows three possible NC scores (0, 1, and 2) and the probabilities of obtaining those scores are $Q_1(\theta_t)Q_2(\theta_t)$, $P_1(\theta_t)Q_2(\theta_t)$, and $P_1(\theta_t)P_2(\theta_t)$, respectively. Then each item is added in turn, adjusting the accumulating probabilities under the conditions that the added item is answered correctly or incorrectly. Once the final *n*th item is added, the probabilities become the distribution of NC scores for the test at the $\theta_t$. This recursive algorithm generalizes readily to items scored in more than two categories (Thissen et al., 1995).

If ability estimates in an MST are computed by the ENC scoring method rather than by MLE, measurement precision of the MST can be projected from recursion-based score distributions rather than by a simulation. This is because the recursive algorithm allows the conditional NC score distributions of modules in the MST at a given ability level to be computed directly. The process of obtaining measurement precision of an MST is briefly described below for a two-stage test followed by a more detailed example with formulas. This generalizes readily to the MST of three stages or more.

(1) Given a specific true ability of $\theta$, compute the observed NC score distribution for a routing module at the first stage using the recursive algorithm.

(2) Divide the NC scores of the routing module to several groups according to RDPs for allocating the next modules at the second stage. For example, if there are three difficulty-level modules at the second stage, the NC scores of the routing module will be classified into three groups based on the RDPs.

(3) Implement a second recursion to compute the NC score distribution for each module at the second stage.

(4) Compute the joint conditional distributions of NC scores of the first and second modules across all allowed routes. These are the distributions of total test NC scores (i.e., summed scores of two modules at the first and the second stage) for the routes.

(5) Map each total test NC score for each route onto the corresponding ability estimate on the IRT scale with the ENC scoring, producing the distribution of ability estimates for each route.

(6) A sum of the distributions of ability estimates across all routes is the distribution of ability estimates of a test given the true ability, $\theta$.

(7) The CSEE at the $\theta$ is then the standard deviation of the distribution of ability estimate for the test. The conditional bias is computed as the difference between the true ability of $\theta$ and the mean of the distribution of ability estimates.

(8) Implement steps 1-7 across a grid of true ability values.

To formulize this procedure with an example of a two-stage MST configuration in the left panel of Figure 1, where the MST has a routing module at the first stage and three difficulty-level modules at the second stage, define $X$ as the NC score and $\theta_t$ as a fixed true ability value. The conditional distribution of NC score $X$ given the $\theta_t$ is obtained using Lord-Wingersky recursion and is defined as:

$$f_{1M}(X \mid \theta = \theta_t). \tag{3.1.7}$$

where $1M$ indicates a routing module at the first stage. In a similar way, the conditional distributions of NC score $X$ for the three modules at the second stage given the $\theta_t$ are defined as:

$$\begin{aligned} &f_{2E}(X \mid \theta = \theta_t), \\ &f_{2M}(X \mid \theta = \theta_t), \\ &f_{2H}(X \mid \theta = \theta_t). \end{aligned} \tag{3.1.8}$$

where 2*E*, 2*M*, and 2*H* denote “easy”, “medium”, and “hard” difficulty level modules at the second stage, respectively.

Now, the conditional distributions of NC scores for three routes (i.e., 1M-2E, 1M-2M, and 1M-2H) are the joint conditional distributions of NC score *X* for the routes. To produce those joint conditional distributions, it should be noted that each module at the second stage can take only certain NC score points from the routing module at the first stage depending on the RDPs. Suppose that a routing module consists of 15 items and each of the three modules at the second stage has 20 items. Also, suppose that two RDPs of NC scores are set to 5 and 11. According to the defined pathways in the left panel of Figure 1, examinees with the NC scores from 0 to 4 will be given the easy module at the second stage, leading to the possible total test NC scores ranging from 0 to 14. Similarly, examinees with the NC scores from 5 to 10 will be allocated the medium module at the second stage, leading to the possible total test NC scores ranging from 5 to 20. Lastly, examinees with the NC scores from 11 to 15 are assigned the hard module at the second stage, leading to the possible total test NC scores ranging from 11 to 25. Note that the probabilities of having total test NC scores outside the possible score range for each route are zero. Following this rule, three joint conditional distributions of total test NC scores for the three routes are given by:

$$
\begin{aligned}
g_{1M-2E}(X \mid \theta = \theta_t) &= \sum_{i \in 1M_{2E}} \sum_{j \in 2E} f_{1M}(X_i \mid \theta = \theta_t) f_{2E}(X_j \mid \theta = \theta_t), \\
g_{1M-2M}(X \mid \theta = \theta_t) &= \sum_{i \in 1M_{2M}} \sum_{j \in 2M} f_{1M}(X_i \mid \theta = \theta_t) f_{2M}(X_j \mid \theta = \theta_t), \\
g_{1M-2H}(X \mid \theta = \theta_t) &= \sum_{i \in 1M_{2H}} \sum_{j \in 2H} f_{1M}(X_i \mid \theta = \theta_t) f_{2H}(X_j \mid \theta = \theta_t).
\end{aligned}
\tag{3.1.9}
$$

where $1M_{2E}$, $1M_{2M}$, and $1M_{2H}$ represent the subsets of NC scores of the routing module to which the easy, medium, and hard modules at the second stage are assigned, respectively. Now, $g_{1M-2E}(X|\theta = \theta_t)$, $g_{1M-2M}(X|\theta = \theta_t)d$, and $g_{1M-2H}(X|\theta = \theta_t)$ in Equation (3.1.9) are the joint conditional distributions of NC scores for the 1M-2E, 1M-2M, and 1M-2H routes. Figure 2 displays the process described through Equations (3.1.7) to Equation (3.1.9) graphically.

Next step is then to map the NC scores for each route to the corresponding IRT ability estimates using the ENC scoring method. The three joint conditional distributions of estimated ability given the $\theta_t$ are defined as:

$$
\begin{aligned}
&g_{1M-2E}(\theta_X \mid \theta = \theta_t), \\
&g_{1M-2M}(\theta_X \mid \theta = \theta_t), \\
&g_{1M-2H}(\theta_X \mid \theta = \theta_t),
\end{aligned}
\tag{3.1.10}
$$

where $\hat{\theta}_X$ denotes the IRT ability estimate corresponding to the NC score of each route. Finally, a conditional distribution of ability estimates of the MST given the $\theta_t$ consists of the three joint conditional distributions and is defined as follows:

$$
h(\theta_X \mid \theta = \theta_t) = \begin{cases} g_{1M-2E}(\theta_X \mid \theta = \theta_t) \\ g_{1M-2M}(\theta_X \mid \theta = \theta_t). \\ g_{1M-2H}(\theta_X \mid \theta = \theta_t) \end{cases}
\tag{3.1.11}
$$

Therefore, the sum of the area under the conditional distribution of $h(\hat{\theta}_X|\theta = \theta_t)$ is a unit value.

A conditional expected value and a variance of the ability estimates under the distribution of $h(\hat{\theta}_X|\theta = \theta_t)$ can be computed as:

$$E(\theta_X \mid \theta = \theta_t) = \sum_p^P \sum_k^K \theta_k h(\theta_k \mid \theta = \theta_t), \tag{3.1.12}$$

$$Var(\theta_X \mid \theta = \theta_t) = E(\theta_X^2 \mid \theta = \theta_t) - E(\theta_X \mid \theta = \theta_t)^2, \tag{3.1.13}$$

where $K$ represents the total number of ability estimates in each route and $P$ indicates the number of total routes in an MST. The CSEE is then the square root of the conditional variance and is defined as:

$$CSEE(\hat{\theta}_X \mid \theta = \theta_t) = \sqrt{Var(\hat{\theta}_X \mid \theta = \theta_t)}, \tag{3.1.14}$$

In addition, a conditional bias of ability estimate given the $\theta_t$ is computed as:

$$Bias = E(\theta_X \mid \theta = \theta_t) - \theta_t. \tag{3.1.15}$$

The analytical process above can be readily generalized to three- or more stage MSTs by recursively applying the steps used for the two stage MST to each of the joint conditional distributions of NC scores after a subsequent module is added. More specifically, each of the joint conditional distributions in Equations (3.1.9) is assumed as the conditional distribution of a routing module and the procedure described above is recursively applied to the joint conditional distributions with subsequent modules at the next stage. For example, suppose that the CSEE of 1-3-3 MST panel (see the right panel of Figure 1) at a true ability of $\theta_t$ needs to be computed. From the first- and the second stages, three joint conditional distributions of NC scores for three routes (i.e., 1M-2E, 1M-2M, and 1M-2H) are produced through Equations from (3.1.7) to (3.1.9).

Now assume that the second stage is the first stage with three routing modules and that each of three joint conditional distributions is the conditional distribution of NC score for each routing module. For the conditional distribution of the 1M-2E route, two possible joint conditional distributions are computed (i.e., 1M-2E-3E and 1M-2E-3M). For the conditional distribution of the 1M-2M route, three possible joint conditional distributions are computed (i.e., 1M-2M-3E, 1M-2M-3M, and 1M-2M-3H). For the conditional distribution of the 1M-2H route, two possible joint conditional distributions are computed (i.e., 1M-2H-3M and 1M-2H-3H). The NC scores of seven routes are mapped to the corresponding IRT ability estimates, resulting in the conditional distribution of ability estimate of the 1-3-3 MST given the $\theta_t$. The standard deviation of conditional distribution is then the CSEE at the $\theta_t$.

Not only analytically computing the conditional distribution of ability estimates is both computationally simpler and more precise than simulating large numbers of test responses, but also many of the performance characteristics of MSTs can be projected directly from the conditional distribution. In the following section, predicting classification accuracy of an MST, which is one of the important measurement properties of credentialing tests, will be introduced by applying the CSEEs obtained from the conditional distribution of ability estimates.

### 3.1.3 Prediction of MST Classification Accuracy

In a license or certification test, one of the primary interests is classification accuracy. Classification accuracy is the proportion of examinees whose classification results from a test are in agreement with the true classification status of examinees.

Unless a test is perfectly reliable, however, misclassifications are inevitable whenever a classification is made based on a test score due to measurement error. If examinees are classified as mastery or non-mastery, there are two types of classification error: false negative and false positive errors. The false negative error rate is defined as the proportion of examinees whose true abilities belong to the mastery status are incorrectly classified as non-mastery status and the false positive error rate is referred to as the proportion of examinees whose true status are the non-mastery level are incorrectly classified as the mastery status.

When an MST is applied in credentialing tests, it is necessary to compute the expected classification accuracy and the classification error rates in order to keep the efficiency of a test or to examine the impact of varying cut score in the credentialing tests. Rudner (2001, 2005) developed a procedure to estimate the classification accuracy in discrete categories under IRT without a simulation. However, this procedure requires the CSEE at a given ability $\theta$. Recently, Park et al. (2017) predicted the classification accuracy using the analytically computed MST information function by applying Rudner (2001, 2005)'s approach. As indicated in Park et al. (2017), it is possible to simply compute the predicted classification accuracy without a simulation if the CSEEs can be analytically obtained.

The procedure to calculate the classification accuracy consists of two steps. To explain the procedure, let $\theta$ be a true ability value and $\hat{\theta}$ be its estimate. Also, let $\theta_c$ represent the cut score for a test. First, calculate the classification error (false positive error or false negative error) for the $\hat{\theta}$ depending on the location of the $\hat{\theta}$. If $\hat{\theta} > \theta_c$ when $\theta < \theta_c$, the error is a false positive and if $\hat{\theta} < \theta_c$ when $\theta > \theta_c$, the error is a false negative. Asymptotically, the ability estimate $\hat{\theta}$ is assumed to follow a normal distribution of $N(\theta, se(\theta))$, where $se(\theta)$ is a CSEE given the $\theta$. The classification error is given by:

$$\begin{aligned} &P(\hat{\theta} > \theta_c \mid \theta) = 1 - \Phi(z), \quad \text{if } \theta < \theta_c \\ &P(\hat{\theta} < \theta_c \mid \theta) = \Phi(z), \quad \text{if } \theta > \theta_c \end{aligned} \tag{3.1.16}$$

where $P(\hat{\theta} > \theta_c|\theta)$ and $P(\hat{\theta} < \theta_c|\theta)$ are the expected false positive and false negative error rates given the $\theta$, respectively. $\Phi(z)$ represents a cumulative density function of the standardized distance *z* between $\theta_c$ and $\theta$, which is calculated as:

$$z = \frac{\theta_c - \theta}{se(\theta)}. \tag{3.1.17}$$

Note that the obtained classification error in Equation (3.1.16) is the conditional error given the true ability of $\theta$.

To compute the marginal classification error rates, the population distribution of examinees needs to be taken into account in the second step. Now, let $h(\theta)$ be the probability density function of the population distribution (a standard normal distribution is assumed in this study). Then, the marginal classification error rates can be expressed as:

$$\begin{aligned} &FP\ error\ rate = \int_{\theta=\theta_c}^{\infty} 1 - \Phi(z) h(\theta) d\theta, \\ &FN\ error\ rate = \int_{-\infty}^{\theta=\theta_c} \Phi(z) h(\theta) d\theta, \end{aligned} \tag{3.1.18}$$

where *FP* and *FN* denote the false positive and the false negative, respectively. The actual calculation of integrating the conditional classification error functions over the population distribution is approximated by replacing the integration with a summation based on a discrete population distribution and a finite number of equally spaced points.

$$\begin{aligned} FP\ error\ rate &= \sum_{\theta_i < \theta_c} P(\theta > \theta_c \mid \theta_i) A(\theta_i), \\ FN\ error\ rate &= \sum_{\theta_i > \theta_c} P(\theta < \theta_c \mid \theta_i) A(\theta_i), \end{aligned} \tag{3.1.19}$$

where $\theta_i$ and $A(\theta_i)$ denote the node and normalized weight of $\theta$ at a quadrature point *i*. Finally, the classification accuracy is simply defined as $1 - (FP\ Error\ Rate + FN\ Error\ Rate)$.

### 3.2 Simulation

To show that the new analytical evaluation method is able to well predict MST performance with respect to measurement precision and classification accuracy, two simulation studies were conducted. In each simulation study, the analytical measurement properties of an MST computed by the new and Park et al.'s (2017) analytical methods and the empirical measurement properties obtained from two MC-based simulation methods using the ENC and MLE scorings, respectively, were compared. For the first simulation, two criteria of measurement precision, which are a conditional bias and CSEE, resulted from the four different methods were examined. Note that the conditional bias cannot be derived from Park et al.'s analytical method because only the CSEEs can be obtained from the MST TIF. For the second simulation, the classification accuracies of an MST through the CSEEs obtained from the four methods were investigated.

### 3.2.1 Design of Simulation

To construct an MST, a bottom-up assembly approach was used because the focus of Study 1 was to examine the performance of the proposed evaluation method. Two fully crossed factors were included for the MST assembly: MST panel configurations (1-

3 and 1-3-3 MSTs) and test lengths (24 and 48 items). The two panel configurations are commonly used in previous research and testing programs (e.g., Jodoin et al., 2006; Luecht et al, 2006; Luecht & Nungester, 1998; Wendler & Bridgeman, 2014). Two levels of length were chosen to represent the short and moderate test length conditions in MST.

For all MST panels items were evenly distributed to each module across stages. For example, twelve items were assigned to the first- and the second-stage modules, respectively, for the 1-3 MST with a test length of 24. In addition, all test forms were built with a content constraint that every module had to consist of four categorical content of items with the same proportions of [.25, .25, .25, .25]. In any assembled MST panels, no overlapped items were allowed between stages, but overlapped items were allowed within a same stage.

### 3.2.2 Item Pool

A simulated item pool with 300 items was used for the two simulation studies. The IRT 3PL model was employed to generate the items based on the item parameter statistics of a large-scale license examination used in Luo and Kim (2018). The $a$-, $b$-, and $c$-parameters were generated from $logN(0.0, 0.3)$, $N(0.0, 1.0)$, and $Beta(5, 42)$, respectively. Table 1 provides a summary of descriptive statistics of the generated item parameters in the item pool. Each of the items in the pool was randomly given one of four content categories and each of four content categories had the same proportion of items in the bank, which were [.25, .25, .25, .25].

### 3.2.3 Module Information Function Target

The procedure used in Luo and Kim (2018) was adapted to develop the MIF targets for the bottom-up assembly approach. Let $n$ be the length of each module in an MST and $I_{avg}(\theta)$ be the average of the *l* largest information of items at $\theta_i$ among the item pool, where *l* denotes the test length. The MIF target of each module was then developed as such $n \times I_{avg}(\theta)$ with five points of $\theta$s, depending on the location of ability level where each module measures examinee proficiency precisely. Assuming the population of examinees follows a standard normal distribution, the ability scale was divided into three intervals with two $\theta$ cutoff points of $-0.44$ and 0.44, which roughly grouped the population into three equal-size subpopulations. Therefore, with a restricted boundary of [$-2.00$, 2.00] on the ability scale, the three ability intervals were [$-2.00$, $-0.44$], [$-0.44$, 0.44], and [0.44, 2.00]. The MIF targets at each stage were then developed to represent each of those ability intervals except the routing module at the first stage. The easy difficulty module should accurately measure examinees' proficiencies in the range of [$-2.00$, $-0.44$], the medium difficulty module needs to cover the proficiency range of [$-0.44$, 0.44], and the hard difficulty module should have optimized test information values in the proficiency range of [0.44, 2.00]. The five $\theta$s for the routing MIF target were set to $-2.00$, $-1.22$, 0.00, 1.22, and 2.00. The three middle points are median values of the three intervals. For the modules at the subsequent stages, the five points were fixed by equally spacing four areas at each of the three intervals. For example, the five $\theta_i$s are $-2.00$, $-1.61$, $-1.22$, 0.83, and $-0.44$ for the interval of [$-2.00$, $-0.44$]. Note that other test constraints (e.g., content specification and exclusion of enemy items) were not considered during the development of the MIF targets because the goal of these targets was to assemble MIFs with high TIFs (Luo & Kim, 2018).

### 3.2.4 MST Assembly

The MIP approach was used to assemble the MST. Given the developed MIF targets, the goal of the assembly was minimizing the distances between the MIF targets and the observed MIFs from the assembled MST over the specified $\theta$s as well as satisfying all test constraints. The mathematical formulas for the ATA model using a bottom-up approach are explained as follows.

In the ATA model items in the pool are denoted as $i = 1, \ldots, I$ and modules assembled in the entire MST panel are represented as $f = 1, \ldots, F$. Now let $V_R$ be a route (i.e., the combinations of modules across all stages) that each examinee will take during the whole process of testing and $n_m$ be the number of items at each module. Also, let $V_c$ denote a subset of items which belong to content category $c$ in the pool and $n_c$ be the number of items that must be included in each module from this subset. $I_i(\theta_k)$ and $T_{\theta_k}$ are used to represent the information of item *i* and the MIF at $\theta_k$, respectively. The objective function of the ATA model is then expressed as:

$$\min y \tag{3.2.1}$$

subject to

$$\sum\nolimits_{i=1}^{I} I_i(\theta_k) x_{fi} \le T_{\theta_k} + y, \quad \text{for all five } \theta_k\text{s and all modules,} \tag{3.2.2}$$

$$\sum\nolimits_{i=1}^{I} I_i(\theta_k) x_{fi} \ge T_{\theta_k} - y, \quad \text{for all five } \theta_k\text{s and all modules,} \tag{3.2.3}$$

$$\sum\nolimits_{f \in V_R} x_{fi} \le 1, \quad \text{for all } p \text{ and all } i, \tag{3.2.4}$$

$$\sum_{i=1}^{I} x_{fi} = n_m, \quad \text{for all modules,} \tag{3.2.5}$$

$$\sum_{i \in V_c} x_{fi} = n_c, \quad \text{for all } c \text{ and all modules,} \tag{3.2.6}$$

$$x_{fi} \in \{0, 1\}, \quad \text{for all } i, \tag{3.2.7}$$

$$y \geq 0. \tag{3.2.8}$$

where $x_{fi}$ is a binary decision variable indicating whether an item is selected for each module. The constraint in Equation (3.2.4) guarantees that no item overlap between the three stages and a fixed length of each module is imposed by Equation (3.2.5). The content category requirements for each module are modeled by the constraint in Equation (3.2.6). For each module and $\theta_k$ value, the distance between the target value $T_{\theta_k}$and the MIF of the assembled form is constrained to be no greater than $y$ by Equation (3.2.2) and (3.2.3).

### 3.2.5 Analysis

Two different simulations were conducted in Study 1. For the first simulation, where measurement precision was computed by the two analytical methods and two MC-based simulation methods, true abilities were generated from $-4.0$ to 4.0 by increments of 0.1. A total of 81 true ability points, therefore, were used. To examine whether the measurement precision resulted from the two MC-based simulation methods become closer to that obtained from the two analytical method as the sample size increases, the MST were replicated 100, 1,000 and 5,000 times at each ability point.

For the second simulation, where the predicted classification accuracies derived through the CSEEs obtained from the four methods were compared, true abilities of

1,000 and 5,000 examinees were randomly drawn $N(0, 1)$ truncated between $-4.0$ and 4.0. The truncated distribution was used to avoid any negative effects due to outliers. Two cutoff scores of 0.0 and 0.524 were used to predict correct classification accuracy of a pass/fail exam. They correspond to 50%, and 70% passing rates, respectively, assuming the standard normal distribution of the population. For the two MC-based simulated methods based on the ENC and MLE scorings, the classification accuracy was simply predicted by calculating the observed proportion of correctly/incorrectly classified examinees relative to their true pass/fail status. In the simulation, 100 replications were performed for each sample size and the average of predicted classification accuracy of the two simulation methods were used in the comparison.

For both simulation studies, the AMI rule was employed for routing examinees to the modules. For the MC-based simulation method based on the ENC scoring, the ENC scoring was used for both interim and final ability estimation. For the simulation method based on the MLE scoring, the MLE was used for the final ability estimation with a restriction of $[-5.0, 5.0]$, the EAP scoring with a standardized normal prior was applied for the interim ability estimation.

To assemble MSTs using the MIP method, the package "lpSolveAPI" (Diao & van der Linden, 2011; Konis, 2009) of R software (R Core Team, 2016) was used, which provides a convenient application programming interface to a free software of lp_solve version 5.5. Each of the MST panels are assembled with a time limit of 60 seconds. The sub-optimal assembly result, which refers to satisfying the objective of the ATA model, that was achieved within the time limit is considered the final assembled MST. All other procedures are conducted with written R code.

## 3.3 Results

A total of four MSTs (i.e., the 1-3-3 MSTs with 24 and 48 items and the 1-3 MSTs with 24 and 48 items) were successfully assembled with full satisfaction of statistical and non-statistical test specifications. For each of the four MSTs, the analytical and empirical measurement properties of MSTs were obtained and compared below.

### 3.3.1 Measurement Precision

#### 3.3.1.1 Conditional Standard Error of Ability Estimates

Figures 3 through 5 present the CSEEs along the $\theta$ scales obtained from four methods – two analytical methods and two MC-based simulation methods – with 100, 1,000, and 5,000 replications at each $\theta$ point, respectively. The simulation results showed that, when the ENC scoring was used, the empirical CSEEs were very close to the CSEEs computed by the new analytical method regardless of MST panel configurations and test lengths. Given other conditions were the same, the empirical CSEEs tended to converge to the analytical CSEEs when the number of replications at each $\theta$ point increased. For example, the analytical and empirical CSEEs based on the ENC scoring were indistinguishable along the $\theta$ scale under the 5,000 replications (see Figure 5). These results imply that the new analytical method assesses measurement precision of the test more exactly than the simulation method because the simulation always included the uncertainty due to the random error.

When the MLE was employed for the final scoring, however, there were relatively large discrepancies in the CSEEs between Park et al.'s (2017) analytical approach and the MC-based simulation method across all simulation conditions. More specifically, Park et al.'s analytical approach resulted in slightly higher CSEEs than the empirical CSEEs at around $-1.5 \le \theta \le 1.5$ for most of conditions. Outside of this range, the empirical CSEEs were fairly higher than the analytical CSEEs. Unlike the empirical CSEEs based on the ENC scoring, the empirical CSEEs based on the MLE did not approach the analytical CSEEs even though the number of replications increased for all simulation conditions. These results conflict with the findings of Park et al. (2017) because they argued that the analytical CSEEs, which are the inverse of the MST TIF, were very similar to the empirical CSEEs across a wide range of $\theta$ scale. Several reasons that might produce the different results are discussed in detail later.

Interestingly, the MC-based simulation method with the MLE yielded the CSEEs very close to those from the new analytical and simulation methods with the ENC scoring at around $-1.5 \le \theta \le 1.5$ where the tests provide the most information regardless of simulation conditions. At low and high $\theta$ levels, however, the empirical CSEEs based on the MLE were significantly higher than the analytical and empirical CSEEs based on the ENC scoring, implying that the ability estimates from the MLE were more variable than those from the ENC scoring at the extreme $\theta$ levels. Looking at the high $\theta$ levels (e.g., $\theta \ge 2.0$), both the analytical and empirical CSEEs based on the ENC scoring were much smaller than those at the low $\theta$ levels (e.g., $\theta \le -2.0$). To find the reason for this observation, the ability estimates obtained from the ENC scoring were closely examined. In all simulation conditions, it was found that most of the perfect NC scores were mapped onto the ability estimates at around 3.0 while all zero NC scores were mapped onto the ability estimate of $-5.0$. This indicates that the ranges of ability estimates for the examinees with the high $\theta$ levels were much shorter than those for the examinees with the low $\theta$ levels. Due to this fact, the conditional distributions of ability estimates at the high $\theta$ levels had relatively smaller variances than those at the low $\theta$ levels, leading to lower CSEEs at the high $\theta$ levels.

In addition, as test length increased from 24 to 48 items when other conditions were controlled, the range of ability where the CSEEs were similar among the four methods increased.

#### 3.3.1.2 Conditional Bias of Ability Estimates

Figures 6 through 8 show the conditional biases obtained from the three methods – the new analytical method and two MC-based simulation methods – with 100, 1,000, and 5,000 replications at each $\theta$ point, respectively. As noted earlier, Park et al.'s (2017) analytical approach is not able to derive the conditional bias of ability estimates. The patterns of results for the conditional bias were similar to those for the CSEE. When it comes to the ENC scoring, the empirical conditional biases were quite close to the analytical conditional biases along the $\theta$ scale under the 100 replications at each θ point regardless of MST panel configurations and test lengths (see Figure 6). When the number of the replications was large (e.g., 5,000 replications), the conditional biases from the analytical and the MC-based simulation methods were almost the same. Again, these results indicate that the new analytical method evaluates the measurement precision of MSTs more exactly than the simulation method.

Regarding the comparison of two scoring methods, the analytic and empirical conditional biases based on the ENC scoring and the empirical conditional biases based on the MLE were close to zero at around $-2.0 \le \theta \le 2.0$ for all simulation conditions.

The more items the tests contained, the larger the range of $\theta$ where the conditional biases obtained from the two scoring methods overlapped. At the high $\theta$ levels, however, the empirical conditional biases based on the MLE had positive values whereas the analytical and empirical conditional biases based on the ENC scoring presented negative values, meaning that the MLE overestimated examinees' abilities while the ENC scoring underestimated the abilities at that area of $\theta$s. Contrary to this result, the MLE consistently underestimated the abilities at the low $\theta$ levels. In the meantime, the ENC scoring underestimated the ability at the low $\theta$ levels but overestimated the abilities at the extremely low θ levels.

### 3.3.2 Classification Accuracy

Tables 2 and 3 display the correct classification rates and total error rates (i.e., the sum of false positive and false negative rates) predicted from the four methods – two analytical methods and two MC-based simulation methods – for the 1-3-3 MSTs and 1-3 MSTs, respectively. Note that the analytical method with the MLE represents Park et al.'s (2017) analytical approach.

In terms of the ENC scoring, the analytical and MC-based simulation methods produced the classification accuracies that were comparable across MST panel configurations, passing rates, and test lengths. Thus, it can be concluded that the analytical and empirical classification accuracies based on the ENC scoring resulted in similar classification accuracies. With respect to the MLE scoring, however, the analytical and empirical results of classification accuracies had relatively large differences. For example, for the 1-3-3 MSTs, the absolute differences in the correct classification rates between the analytical and simulation methods with the ENC scoring ranged from 0.07 to 0.41 while those between the analytical and simulation methods with the MLE ranged from 1.50 to 2.24 (see Table 2). Similar results regarding the absolute differences were observed for the 1-3 MSTs (see Table 3). No clear pattern was observed regarding the effects of sample size, test length, and the passing rate on the results of classification accuracy regardless of scoring methods.

## 3.4 Discussion and Conclusions

Study 1 proposed a new analytical method to evaluate the performance of an MST, which is based on the ENC scoring for ability estimation. The new analytical method computes the measurement precision of an MST (i.e., the conditional bias and CSEE) from the conditional distribution of ability estimates using the recursive algorithm (Lord & Wingersky, 1984). Once the analytical CSEEs are derived, the classification accuracy of the MST can be estimated without a simulation by applying Rudner’s approach (2001, 2005).

When evaluating the performance of an MST, conducting MC-based simulations usually may cost a significant amount of time, effort, and computing resources depending on the number of simulation conditions (e.g., MST panel designs and test length) and other factors. For example, as more simulation conditions are investigated, the amount of computer storage required to save the simulation data and the time to simulate the tests rapidly increase. On the contrary, analytically deriving the conditional distribution of the ability estimates is computationally simpler and faster than conducting thousands of replications at a grid of the $\theta$ scale. The new analytical method just requires the item parameters, the conversion table where the NC scores of all routes in the MST are mapped onto the IRT scaled ability estimates, and the RDPs. Using the information, it takes only few seconds to calculate the performance characteristics of the MST. Even for the multiple MST designs, the computation will not take more than a minute. In addition, the simulation results in the study showed that the new analytical method produced more exact measurement precision of an MST than the MC-based simulation method. It was shown that, when the ENC scoring was used, the empirical conditional biases and CSEEs along the $\theta$ scale converged to the analytical conditional biases and CSEEs as the number of replications at each $\theta$ point increased. This is because the analytical method is able to compute the measurement precision without relying on the number of examinees at each $\theta$ value. Accordingly, the new analytical method provides researchers a more efficient way of evaluating the measurement properties of an MST.

Although Park el al. (2017) argued that the analytical method based on the MST test information predicted the MST performances as closely as the MC-based simulation method did, the simulation results in this study disclosed that their analytical approach may not be generalized to other situations. Regarding the reasons that might cause the large discrepancies in the CSEEs between Park et al.'s analytical and the simulation method with the MLE, there are external and internal possible factors. First, the external factor is related to the difference of simulation designs between the two studies. In Park et al. (2017), they replicated 100 times of simulation tests using 1,000 examinees generated from $N(0,1)$. In each replication, they analytically computed the standard errors of ability estimates for examinees and then the standard errors were averaged over the replications. Then, CSEEs were calculated on thirteen $\theta$ points, ranging from $-3.5$ to

3.5 in increments of 0.5. For example, the standard errors between −3.25 to −2.75 were averaged to obtain the CSEE at $\theta = -3.0$. In this study, however, a large number (i.e., 100, 1,000, and 5,000) of tests were replicated at each of $\theta$ points, ranging from −4.0 to 4.0 with increments of 0.1, to produce the empirical conditional distribution of ability estimates. Then, the standard deviation of the distribution was the CSEE at each $\theta$ point. In addition, Park et al. used the mixed-format tests with the dichotomous and polytomous items and the EAP estimation for the final scoring whereas the simulations in this study was carried out with the single-format tests with dichotomous items and MLE scoring.

Second, the internal factors have to do with Park et al.'s (2017) analytical approach itself. Their analytical method does not take the exact TIFs of the secondary routes into consideration and calculates the simple moving average using 5 points around each $\theta$ without a specific reason when deriving the MST TIF. Thus, it is likely that the MST TIF differs depending on the shapes of TIFs for the secondary routes and the location and/or number of the points being used to compute the simple moving average. In fact, as noted in Chapter 2, a preliminary simulation showed that the MST TIF varied by the location and/or number $\theta$ points. Consequently, both the external and internal possible factors indirectly support that the new analytical method would be a better alternative for evaluating the MST performance than Park et al.'s analytical method since the proposed method produced more stable and credible measurement precision and accuracies of MST regardless of the simulation conditions.

One may be concerned about the use of the ENC scoring for ability estimation because it is known that the IRT scaled summed score-based scorings might not be precise as the IRT pattern-based scoring methods such as MLE. It was found from the simulation results in the study, however, that the ENC scoring yielded measurement properties very close to those of the MLE scoring at around $-1.5 \le \theta \le 1.5$, where the tests showed the most information and the majority of examinees (i.e., about 87%) exists if the population distribution follows $N(0,1)$. This result is consistent with findings of previous studies (e.g., Kolen & Tong, 2010; Stocking, 1996) that the IRT scaled summed score-based scorings had comparable measurement accuracy and efficiency with the IRT pattern-based scorings. For example, Stocking (1996) showed that the ENC scoring method could be a feasible alternative to the MLE under the adaptive testing context despite a small loss of information. Besides the acceptable measurement precision, the ENC scoring methods has other practical advantages. For example, this scoring method is more understandable for test takers than the MLE when the scoring report is released to public (Thissen et al., 1995). Also, the ability estimation with the ENC scoring method is more robust to suboptimal test adaptation in CAT or misleading responses due to nuisance factors, such as misunderstanding the directions, anxiety, and poor time management (Meijer & Nering, 1997; Stoking, 1996; Stocking, Steffen, & Eignor, 2002).

There were a few limitations in this study. First, the new analytical method was applied to only few MST designs and single-format tests in the simulation studies. However, the proposed method can be easily generalized to more complicated MST designs and mixed-format tests. Even with more stages and modules per a stage, the conditional distribution of ability estimates can be simply computed using the recursive algorithm. The recursion formula also can be readily generalized to the polytomous response data (Thissen et al., 1995). Second, although the ENC scoring has acceptable measurement precision and several practical benefits in ability estimation, it still has some disadvantages. One of them is that for a test with small number of items, the ability estimation of the ENC scoring might not be as accurate as that of the IRT pattern-based scorings. But, the short tests with 24 items in this study resulted in quite similar measurement precision of ability estimation between the ENC and MLE scoring methods. Because many testing programs often use more than 20 items in a test, the ENC scoring would not have severe negative effects on the ability estimation as long as a reasonable number of items are administered to examinees in an MST. Another disadvantage is that a small value of 0.5 was added and subtracted from the zero and perfect NC scores and the possible ability estimates were restricted within $-5.0 \le \theta \le 5.0$ when the ENC scoring was employed. Thus, low and high levels of ability estimates will differ by a small amount and the range of possible ability estimates, which, consequently, affects the measurement accuracy of low and high levels of abilities. Furthermore, when a test contains items following the IRT 3PL model, ability estimates of the NC scores less than the sum of *c*-parameters were obtained by applying linear interpolation. This might produce inaccurate ability estimates for low ability level examinees. In future studies, the effects of these factors on the accuracy of ability estimation need to be investigated. The good news is, however, that these factors have nothing to do with the performance of the new analytical method itself, meaning that if these factors cause less precise ability estimates at the low and high ability levels, the proposed method will return the measurement precision by exactly reflecting the amount of inaccuracy at those levels.

To conclude, the new analytical method would provide an efficient tool to evaluate the measurement performance of an MST. It is expected that this method especially will show its competence in case that many MST designs need to be compared to find a design that has the best measurement performance. Even if the MLE should be used for the final scoring, the proposed method could be a useful alternative of MC-based simulation method because it approximates the CSEEs of the MLE scoring at a wide range of $\theta$ scale.

**Table 1.** Descriptive Statistics of Item Parameters of the Item Pool in Study 1

| Parameter | Mean | SD | Min | Max |
|---|---|---|---|---|
| a | 1.02 | 0.29 | 0.51 | 2.08 |
| b | 0.04 | 0.99 | -2.39 | 2.55 |
| c | 0.10 | 0.05 | 0.02 | 0.28 |

**Table 2.** Classification Accuracies of the Two Analytical Methods and Two MC-based Simulation Methods for the 1-3-3 MST

| | | | Analytical method | | Simulation ($N = 1{,}000$) | | Simulation ($N = 5{,}000$) | |
|---|---|---|---|---|---|---|---|---|
| Test length | Scoring | Passing rate (%) | CCR | TER | CCR | TER | CCR | TER |
| 24 | ENC | 50 | 91.70 | 8.30 | 91.63 | 8.38 | 92.05 | 7.95 |
| | | 70 | 92.67 | 7.33 | 93.08 | 6.92 | 92.81 | 7.19 |
| | MLE | 50 | 89.97 | 10.03 | 91.72 | 8.29 | 92.21 | 7.79 |
| | | 70 | 91.15 | 8.85 | 93.12 | 6.88 | 92.99 | 7.01 |
| 48 | ENC | 50 | 93.79 | 6.21 | 93.64 | 6.36 | 94.14 | 5.86 |
| | | 70 | 94.87 | 5.13 | 94.98 | 5.02 | 94.75 | 5.25 |
| | MLE | 50 | 92.44 | 7.56 | 93.93 | 6.07 | 94.35 | 5.65 |
| | | 70 | 93.37 | 6.63 | 95.27 | 4.73 | 94.97 | 5.03 |

*Note*. ENC = equated number-correct scoring; MLE = maximum likelihood estimation; CCR = correct classification rate; TER = total error rate; *N* = sample size.

**Table 3.** Classification Accuracies of the Two Analytical Methods and Two MC-based Simulation Methods for the 1-3 MST

| | | | Analytical method | | Simulation ($N = 1{,}000$) | | Simulation ($N = 5{,}000$) | |
|---|---|---|---|---|---|---|---|---|
| Test length | Scoring | Passing rate (%) | CCR | TER | CCR | TER | CCR | TER |
| 24 | ENC | 50 | 91.37 | 8.63 | 90.95 | 9.05 | 91.44 | 8.56 |
| | | 70 | 92.65 | 7.35 | 93.02 | 6.98 | 92.76 | 7.24 |
| | MLE | 50 | 90.67 | 9.33 | 91.62 | 8.38 | 92.11 | 7.89 |
| | | 70 | 91.83 | 8.17 | 93.20 | 6.81 | 92.91 | 7.09 |
| 48 | ENC | 50 | 93.80 | 6.20 | 93.67 | 6.33 | 94.13 | 5.87 |
| | | 70 | 94.59 | 5.41 | 94.74 | 5.26 | 94.48 | 5.52 |
| | MLE | 50 | 92.93 | 7.07 | 94.04 | 5.96 | 94.37 | 5.63 |
| | | 70 | 93.91 | 6.09 | 95.06 | 4.94 | 94.74 | 5.26 |

*Note*. ENC = equated number-correct scoring; MLE = maximum likelihood estimation; CCR = correct classification rate; TER = total error rate; *N* = sample size.

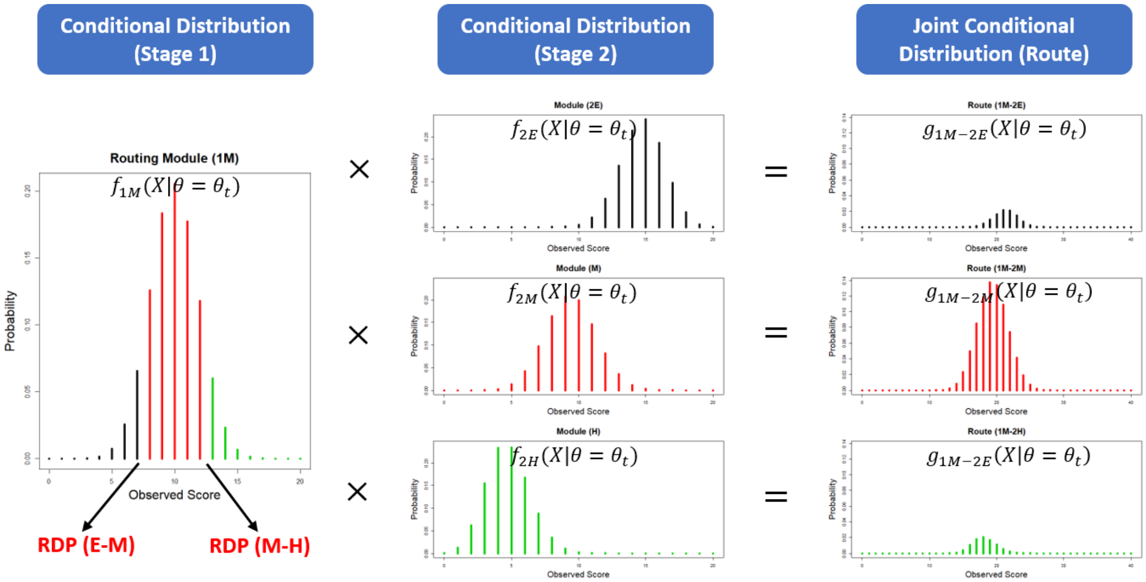

**Figure 2.** A process of computing joint conditions distributions of number-correct scores for the 1-3 MST

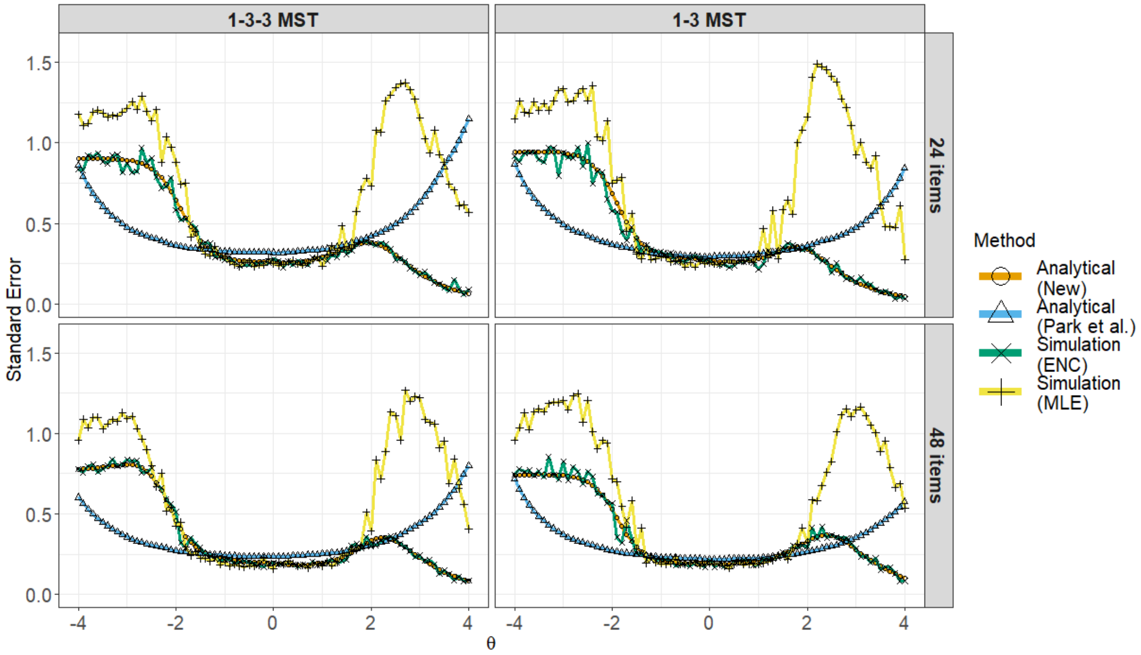

**Figure 3.** Conditional standard errors of ability estimates for the two analytical methods and two MC-based simulation methods with 100 replications

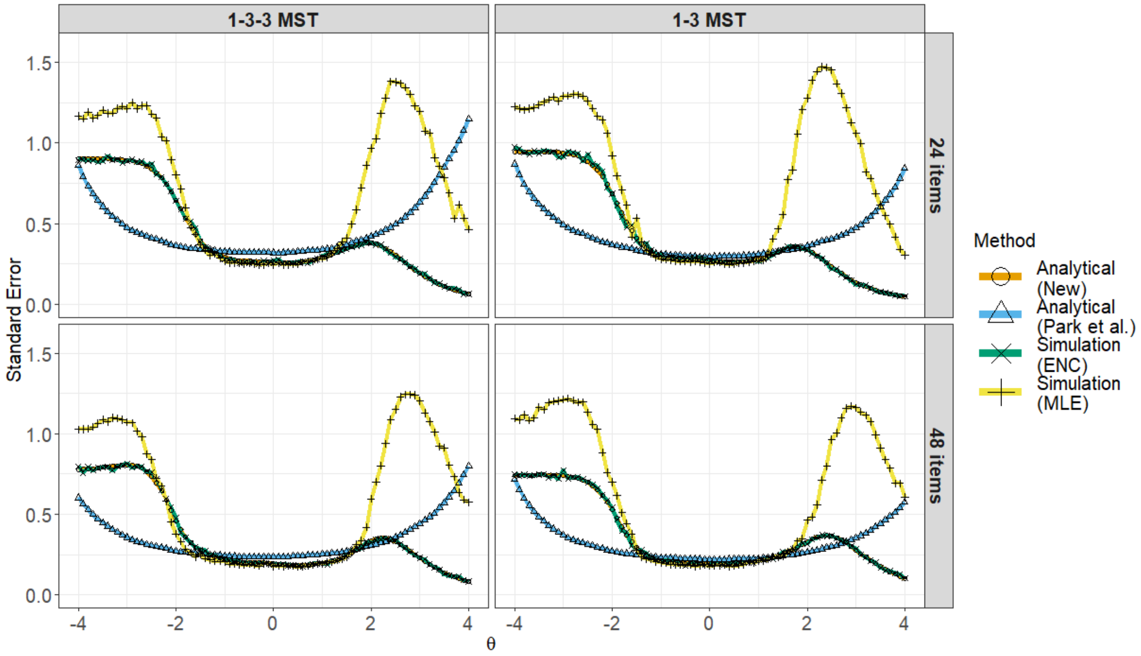

**Figure 4.** Conditional standard errors of ability estimates for the two analytical methods and two MC-based simulation methods with 1,000 replications

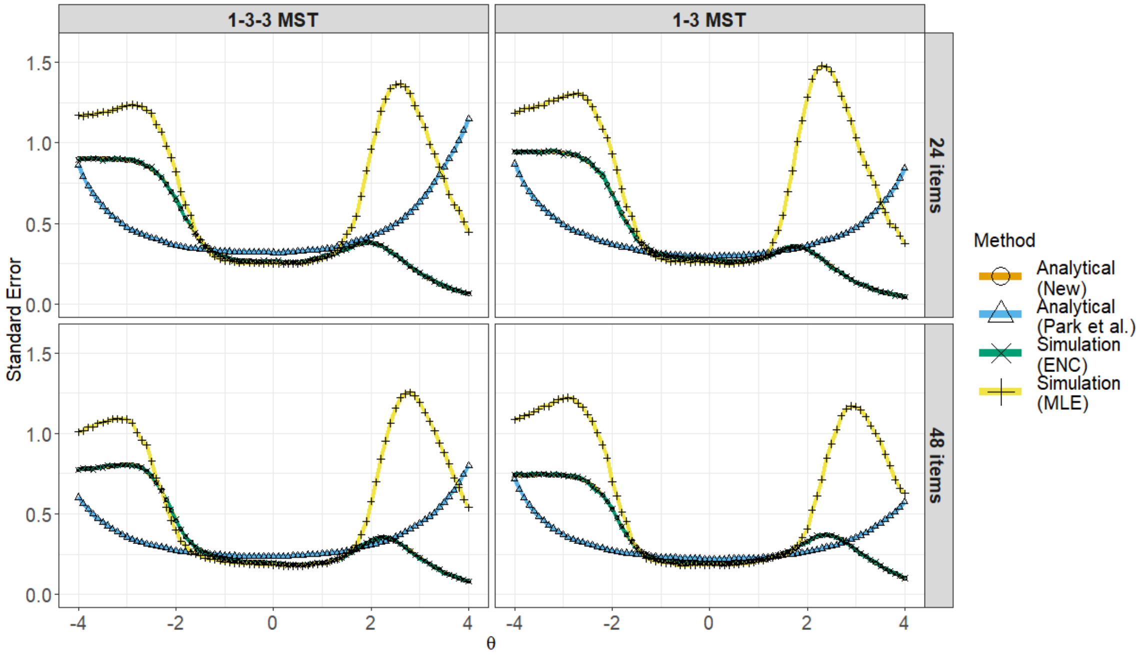

**Figure 5.** Conditional standard errors of ability estimates for the two analytical methods and two MC-based simulation methods with 5,000 replications

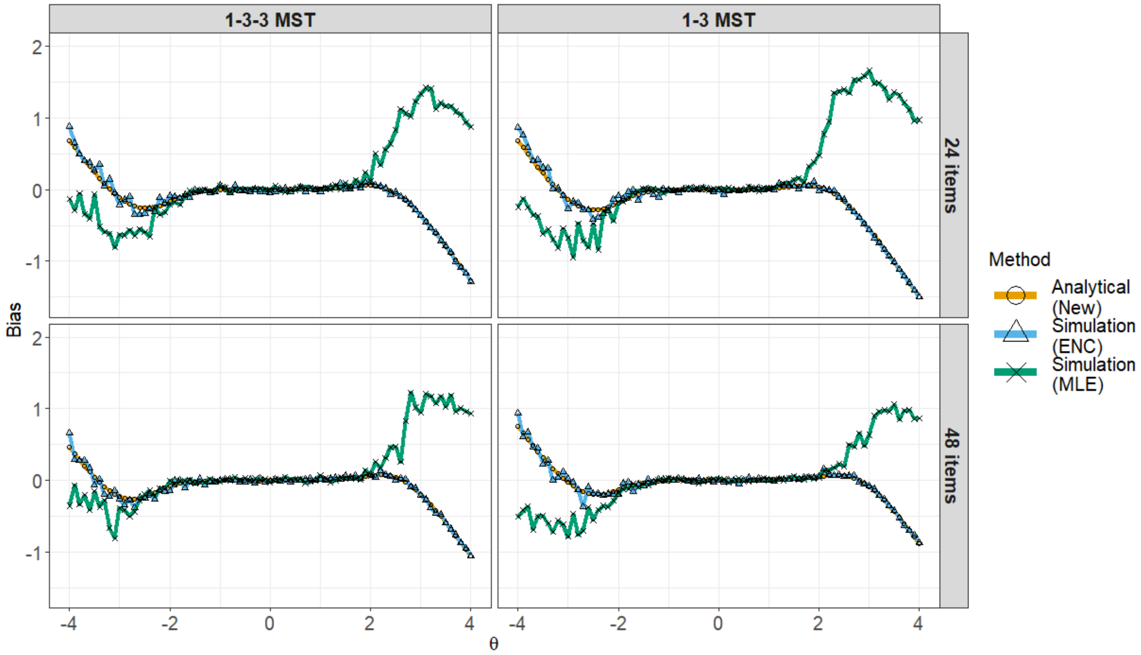

**Figure 6.** Conditional biases of ability estimates for the new analytical method and two MC-based simulation methods with 100 replications

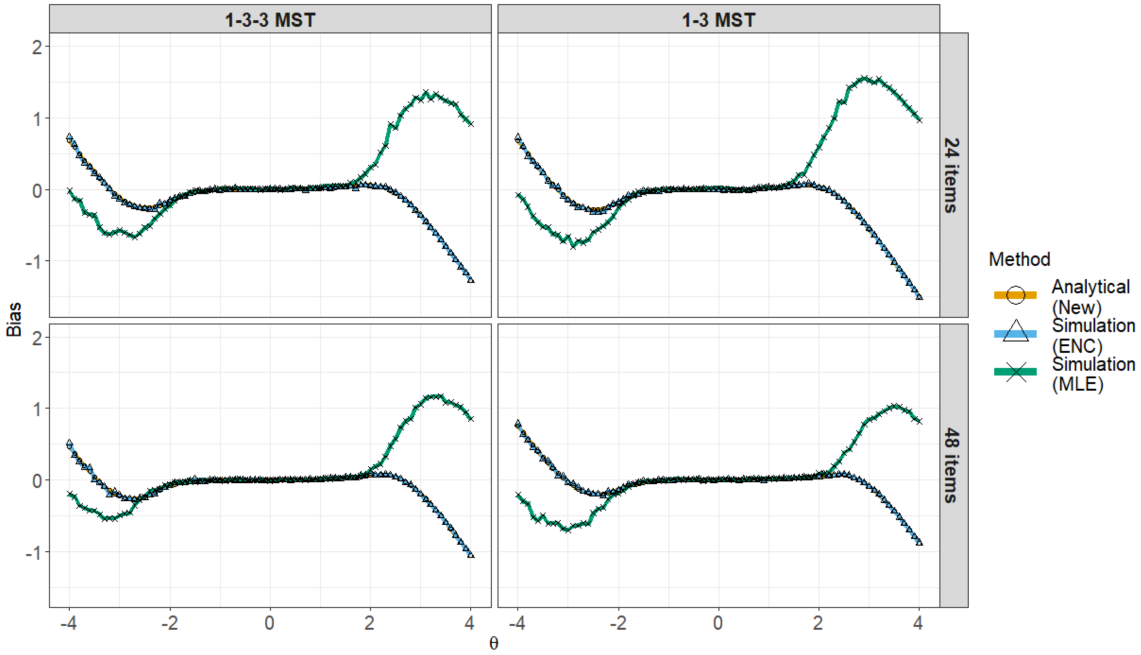

**Figure 7.** Conditional biases of ability estimates for the new analytical method and two MC-based simulation methods with 1,000 replications

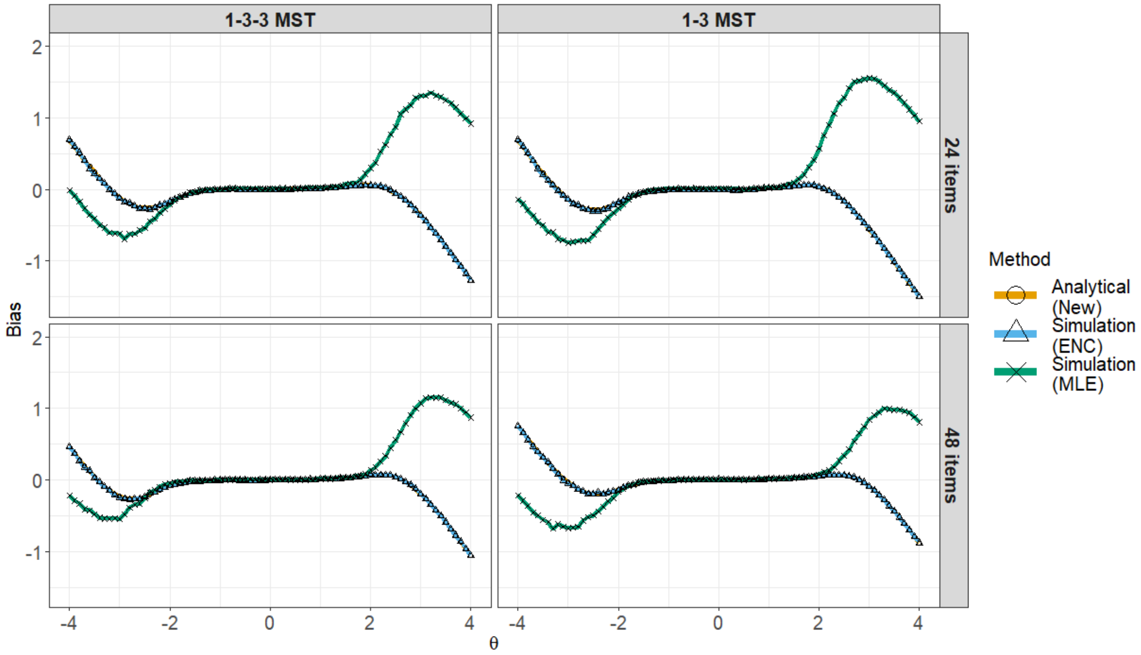

**Figure 8.** Conditional biases of ability estimates for the new analytical method and two MC-based simulation methods with 5,000 replications

# CHAPTER 4

## STUDY 2

The purpose of Study 2 is to propose a process for creating a specific MST design that shows optimal measurement properties. As introdued in Chapter 2, there are many interrelated design factors, hereafter referred to as the design space, that affect measurement properties of an MST. Even if the scope of the design space is limited to four design considerations of the MST panel described in Chapter 2 (i.e., shape of panel structure, test length, characteristics of module, and item bank and examinee population), it is not feasible to evaluate measurement properties of all possible combinations of those design variables.

One practical approach to finding the optimal MST design that provides the best measurement properties is to restrict the design space being examined. In other words, values which MST panel design variables would take are limited to a reasonable range, depending on a testing context. For example, it might be test length, especially when we know that a certain range in test length is sufficient to produce reasonable precision. In this case, it seems reasonable to evaluate the measurement performance of an MST varying test length within a restricted range.

Even with a restricted range of design space for some variables, however, the number of possible combinations of the MST panel design variables would be still too large to evaluate their measurement properties. Suppose that an optimal MST design needs to be created to develop a credentialing test by considering three MST panel configurations, two test lengths, three module lengths per stage, three different RPDs, and two cutoff scores. In this case, the number of possible combinations for those design variables is 108, which will require an immense amount of time and effort to examine all possible conditions. In other testing situations, we may need to examine a wider range of the MST design space. Therefore, more efficient strategies are required to take a broad range of design space into consideration when discovering the optimal design of an MST.

In most research comparing different MST designs (e.g., Jodoin et al., 2006; Luo & Kim, 2018; Wang et al., 2012; Zheng et al., 2012), the studies are composed of three phases: (1) assembling a test given statistical and non-statistical constraints, (2) repeating the test assembly varying conditions of MST design variables, and (3) evaluating the mesurement properties for each of assembled tests using MC-based simulations. If more efficient strategies are used for each of these three phases, it would be possible to assess the performance of MSTs more efficiently and quickly and, therefore, lead to the examination of a broader range of design space of an MST. In addition, the more design conditions that are evaluated, the more possibly optimal the discovered MST design is.

At the first phase, a top-down assembly approach is applied in the study. As described in Chapter 2, the top-down approach enables a designing process of an MST to be simplified compared to a bottom-up approach because it automatically identifies an optimal partition of test-level variables and other specifications into modules as well as satisfies optimal measurement precision (Luo & Kim, 2018). Accordingly, the use of the top-down approach would be more advantageous than the bottom-up approach as the number of modules and stages become large in the MST panel in that it rapidly increases the number of design parameters that should be considered at the module level.

At the second phase, the design space of an MST is systematically searched, seeking the combination of design variables that produce optimal measurement performance in some sense. This search involves iteratively using an ATA process based on the top-down approach varying the design variables. Among other design variables, varying targeted subpopulations of routes is an essential part in the proposed process. For the third phase, it is necessary to evaluate the performance of an MST quickly and precisely. As demonstrated in Study 1, the new analytical method to compute the CSEEs of an MST is more exact and faster than MC-based simulation methods, both are important advantages given a number of design variants being evaluated. The new analytical method, therefore, is employed in this study.

In fact, the first phase, which is assembling MSTs using a top-down approach, can be considered a part of systematical search for the optimal design of an MST because the iterative process of test assembly at the second phase is based on the top-down approach. Therefore, this study proposes a process of finding an optimally designed MST, given a specific set of testing circumstances, by systematically searching the design space of an MST with the top-down approach and assessing the MST performance with the new analytical evaluation method. In the study, the suggested process consists of four features and each of the features is explained in detail in the following section. A study on the application of the suggested procedure with a real item pool from a large-scale assessment was then conducted to show that the process of finding an optimal MST design is practical and works well.

## 4.1 Process to Find an Optimal MST Design

The process for discovering an optimal MST design has four important features: (1) setting a testing circumstance and MST design space, (2) systematically searching the

MST design space using a top-down approach, (3) analytically evaluating measurement performance of an MST, and (4) computing objective functions. Each of those features is described in detail in the following sub-sections.

### 4.1.1 Setting a Testing Circumstance and MST Design Space

As with any test development, a process for finding an optimal MST design begins with a specific set of test circumstances, most notably a calibrated item pool, content requirements for a test assembly, and an examinee proficiency distribution. Because the circumstances under which MSTs are applied vary case by case, the number and nature of the items available in the pool, the rigor of the content requirements and the location and scale of the examinee proficiency distribution are all factors that dictate whether a given MST design will work well. The purpose of a test is also an important factor that affects the MST design and its performance since test information targets for modules (e.g., when a bottom-up approach is applied) or routes (e.g., when a top-down approach is applied) and the form of the MST panel configuration will be set in response to the testing purpose.

Given a particular testing circumstance, the next step is to determine a scope of design space of the MST panel being searched. Understanding a specific testing context will help restrict the range of design space. For example, a size of the item pool could regulate an appropriate range of test length and the number of parallel MST panels. The statistical characteristics of items (e.g., the distribution of item difficulties) in the pool could constrain the statistical level of differentiation among the modules at each stage as well as the statistical characteristics of the pathways. Also, if the goal of a test is to classify examinees into one of two categories (e.g., pass/fail), then the target module or route containing a cut score should be adapted so that the MST design can result in maximized measurement precision at the cut score.

### 4.1.2 Systematically Searching MST Design Space using a Top-Down approach

#### 4.1.2.1 Assembly of MSTs using a Top-Down Approach

A top-down approach simplifies the designing process of an MST since it is not necessary to set precise specifications at the module level (e.g., module length at each stage and statistical characteristics of MIFs). In the ATA process using the top-down approach, the computer algorithm attempts to find the best decomposition of the test level constraints into the module levels, which, therefore, results in the optimal measurement precision (Luo & Kim, 2018). Though a bottom-up approach can be used to construct the MST, the problem is that there exist too many design parameters at the module level being considered at the ATA process. Searching every single combination of the design variables with the bottom-up approach, therefore, could result in a more time-consuming and laborious effort than the use of the top-down approach. Accordingly, the top-down approach would provide a more efficient way of searching the optimal design of an MST.

In this study, a top-down approach suggested by Luo and Kim (2018) was adapted to assemble the MST. Luo and Kim’s approach features the RIF to constrain statistical specifications at the test level. The following two sub-sections discuss how the RIF objectives are set in this study and the ATA algorithm under the top-down approach.

#### 4.1.2.1.1 Route Information Function (RIF) Objective

The top-down approach for an MST assembly suggested by Luo and Kim (2018) is based on the RIF. A route is referred to as the combination of modules across all stages that an examinee will take to finish a test. Luo and Kim (2018) identified all allowed routes in an MST panel and mapped them with targeted subpopulations. Suppose that a 1-3-3 MST configuration has been selected, and therefore, an examinee will take one of seven possible routes (see the right panel of Figure 1). Since there are three primary routes (i.e., 1M-2E-3E, 1M-2M-3M, and 1M-2H-3H), the examinee population can be divided into three subpopulations according to proficiency levels: low-, middle-, and high-proficiency levels. If the population is assumed to follow a standard normal distribution, the population can be grouped into three equal-size subpopulations of $[-\infty, -0.44]$, $[-0.44, 0.44]$, and $[-0.44, \infty]$ given two RDPs of $-0.44$ and 0.44. Luo and Kim assumed that all allowed routes were mapped onto one of the three targeted subpopulations depending on the selected module at the last stage. For instance, the routes of 1M-2E-3E and 1M-2M-3E should be representative of low-proficiency subpopulation regardless of the selected modules at the second stage. In the ATA process, the objectives are set so that each route has the maximized test information over the ability $\theta$ region of the corresponding targeted subpopulation. Figure 9 illustrates the route mapping to the three targeted subpopulations with the 1-3-3 MST based on Luo and Kim' (2018) strategy.

It seems reasonable that the primary routes should represent the $\theta$ intervals of the targeted subpopulations as in Luo and Kim (2018). It would be more appropriate, however, to assume that the secondary routes (i.e., 1M-2M-3E, 1M-2E-3M, 1M-2M-3E, and 1M-2H-3M) should have the best measurement accuracy at the $\theta$ interval somewhere around the RDPs rather than that they should mapped with only one of the three targeted subpopulations defined by the primary routes. This is because examinees who took the secondary routes are highly likely to have proficiencies around the RDPs (Park et al., 2017). Following this logic, Luo and Kim (2018)'s strategy for the route mapping was modified in this study so that the secondary routes could have the maximized RIFs around the RPDs. A specific method to obtain the $\theta$ intervals of targeted subpopulations for the routes and set the objectives of RIFs is explained below with an example of the 1-3-3 MST (see the right panel of Figure 1).

Recall that the 1-3-3 MST has four secondary routes and Luo and Kim (2018) divided a population into three targeted subpopulations of $[-\infty, -0.44]$, $[-0.44, 0.44]$, and $[-0.44, \infty]$ by the RDPs of $-0.44$ and 0.44. First, substitute $\theta = \pm\infty$ in the $\theta$ intervals of the right and left end subpopulations with finite values (e.g., $\theta = \pm 2$). Then, the three primary routes (i.e., 1M-2E-3E, 1M-2M-3M, and 1M-2H-3H) have their targeted subpopulations at $[-2.00, -0.44]$, $[-0.44, 0.44]$, and $[-0.44, 2.00]$, respectively. Second, it is then expected that the area around the RDPs would be the targeted subpopulations for the secondary routes. Unlike the primary routes, however, no specific interval of the subpopulation is defined for the secondary route. For example, the secondary routes of 1M-2M-3E and 1M-2E-3M are given the subpopulation around the RDP of $-0.44$ and the secondary routes of 1M-2M-3H and 1M-2H-3M are assigned the subpopulation around the RDP of 0.44. Figure 10 displays an illustration of the modified route mapping strategy based on the 1-3-3 MST. Note that since no secondary route exists in the two-stage MST, the modified procedure is not necessary.

Once the targeted subpopulations for all allowed routes are determined, the next step is to set objectives of RIFs for the ATA model. How to set the objectives is aligned with the purpose of a test. In this study, it was assumed that MSTs are intended to measure achievement or growth of students. To attain this purpose, the MSTs should be assembled to have accurate measurement precision over a wide range of the proficiency scale.

When constructing the MST using the bottom-up approach, the target TIFs of the modules are usually developed so that the objectives in the ATA process is to minimize the discrepancies between the target MIFs and the assembled MIFs. Instead of deriving the target RIFs, however, Luo and Kim (2018) used the relative target method where the objectives of each route were to maximize the route information over the $\theta$ interval of the targeted subpopulation. This target strategy is useful when the objective of a test is to have better information along the $\theta$ scale such as broad-range diagnostic testing (van der Linden, 2005). Therefore, the relative target method was applied in this study.

For the relative target in the ATA process, the objective of each route in the MST was set at only one $\theta$ point in this study. Specifically, each of the primary routes had the objective at the mid-point of the $\theta$ interval of the targeted subpopulation and each of the secondary routes had the objective at the corresponding RDP. For example, if three primary routes 1M-2E-3E, 1M-2M-3M, and 1M-2H-3H in the 1-3-3 MST have the targeted subpopulations at $[-2.00, -0.44]$, $[-0.44, 0.44]$, and $[-0.44, 2.00]$, the objective of their RIFs are set at $-1.22$, 0.0, and 1.22. Then, the objective of the secondary RIFs of 1M-2M-3E and 1M-2E-3M are set at the RDP of $-0.44$ and the objective of the secondary RIFs of 1M-2M-3H and 1M-2H-3M are set at the RDP of 0.44. In fact, in Luo and Kim (2018), three $\theta$ points within the $\theta$ interval were used to represent the corresponding targeted supopulation of each route. However, a preliminarly simulation study showed that generally the assembled routes with the objective set at one $\theta$ point had higher TIFs over the $\theta$ intervals of the targeted subpopulations than those with the objective set at the three $\theta$ points.

#### 4.1.2.1.2 Top-Down Approach in ATA

Under the top-down approach, an ATA algorithm with the MIP was used to build MSTs. Given the targeted subpopulations for all allowed routes, the goal of the ATA model was to maximize the TIFs of the assembled routes over the specified $\theta$s, considering other test constraints. The mathematical formulas of the MIP for the top-down assembly of MSTs are explained below.

Suppose that the items in the item pool are denoted as $i = 1, \ldots, I$ and the assembled modules in the MST are indexed as $f = 1, \ldots, F$. Now let $V_R$ be a route (i.e., the combinations of modules across all stages) that an examinee will have during the testing and $n$ represents test length. Also, let $V_c$ be a subset of items in the item pool that are classified into category $c$, $n_c$ represent the number of items from this subset, and $I_i(\theta_k)$ represent the information of item $i$ at $\theta_k$. To maximize the test information over a $\theta$ interval, the object function of the ATA problem is expressed as follows.

$$\max y \tag{4.2.1}$$

subject to

$$\sum_{f \in V_R} \sum_{i} I_i(\theta_k) x_{fi} \geq y, \tag{4.2.2}$$

$$\sum_{f \in V_R} \sum_{i} c_i x_{fi} = n_c, \tag{4.2.3}$$

$$\sum_{f \in V_R} \sum_{i} x_{fi} = n, \tag{4.2.4}$$

$$\sum_{f \in V_R} x_{fi} \le 1, \tag{4.2.5}$$

$$\sum_{i} I_i(\theta_k) x_{f_u i} = \sum_{i} I_i(\theta_k) x_{f_t i}, \tag{4.2.6}$$

$$\sum_{i} x_{fi} \ge n_{min}, \tag{4.2.7}$$

$$x_{fi} = \{0, 1\}, \tag{4.2.8}$$

where $x_{fi}$ denotes a binary decision variable of item *i* for module *f*. If an item is selected in the assembled module, the value of the binary decision variable is 1 otherwise 0. The categorical test specification for each route is constrained by Equation (4.2.3) and the test length of the route is imposed by Equation (4.2.4). No overlap of the items across stages in the MST are modeled in Equation (4.2.5). There are two more important constraints in the ATA model. Equation (4.2.6) guarantees that two adjacent modules $f_u$ and $f_t$ within the same stage intersect at the RPD. Equation (4.2.7) is used to avoid an empty module at any stage, and thus leads to each module having a minimum number of items $n_{min}$.

#### 4.1.2.2 Systematical Search of Design Space

To find the combination of design variables that results in measurement performance that is optimal in some sense, the parameterized design space of an MST was systematically searched. In fact, the systematic search consists of two parts: (1) iteratively assembling MSTs based on a top-down ATA process varying the design parameters and (2) assessing the measurement performance of the series of the assembled MSTs. This subsection focuses on the iterative assembly of MSTs and the evaluation of MST performance is explained next.

Once some MST design variables are restricted or fixed (e.g., test length and the shape of MST panel configuration), the key of searching the design space is iteratively varying the RDPs. This is because the location of the RDPs under the top-down approach is an important factor that affects the measurement performance of an MST. Varying the RPDs determines the targeted subpopuluations of all allowed routes of an MST. Then, the range and location of the targeted subpopulations impact the characteristics of the modules (e.g., length and statistical properties of modules) in the assembled MST through the ATA process. The characteristics of modules will affect the shapes of the MIFs and/or RIFs, which regulate the measurement precision and other critical psychometric properties of the assembled MST.

In this study, all possible combinations of RDPs were searched, given that other design variables were fixed, by systematically varying the locations of RPDs. For example, when the 1-3-3 MST is used, the RDP between low- and mid-proficiency levels is allowed to shift from −0.8 to 0.1 by an increment of 0.02 and the RDP between mid- and high-proficiency levels is allowed to vary from 0.1 to 0.8 by an increment of 0.02. In this case, a total 1,296 combinations should be searched. When the 1-2-2 MST is selected, however, only one RDP is required in that configuration. In this case, the RDP may move from −0.7 to 0.7 by an increment of 0.02. Then, a total of 71 MSTs can be designed given that other design variables are controlled. Recall that once the targeted subpopulations of the primary routes are determined by the RDPs, the subpopulations of the secondary routes are automatically set under the top-down ATA approach.

The wider the range of the proficiency and the smaller incremental value used in the iterative search, the more thoroughly an optimal MST design can be sought. But, it costs more time to assemble the series of MSTs. In fact, it is unnecessary to use too broad a range for the RDPs, meaning that the RDPs should vary within a reasonable $\theta$ interval. For example, for the 1-3-3 MST it may not be appropriate to have an RDP between low- and mid-proficiency levels that is located above $\theta = 0$. Thus, a reasonable range of RDPs and size of increment value should be determined in advance.

### 4.1.3 Analytically Evaluating Measurement Performance of an MST

To find an optimally designed MST, measurement performance of the series of assembled MSTs should be evaluated. In the evaluation process, the CSEEs were computed using the proposed analytical method in Study 1. Recall that examinees' abilities should be estimated using the ENC scoring method to use this method. Based on this scoring method, score distributions and conditional standard errors of ability estimates can be calculated exactly across a grid of ability values using a recursive algorithm (Lord & Wingersky, 1984). As shown in Study 1, the analytical method produces more exact CSEEs and the computation is much faster than performing a simulation.

### 4.1.4 Computing Objective Functions

To find an MST design that provides optimal measurement performance, we need to define the objective function that will serve as the measure of optimality. The objective functions measuring design optimality are computed based on the CSEEs. Several objective functions were suggested in this study, each defining measurement optimality in a different way.

First, marginal test reliability is one possible objective function, which prioritizes the measurement precision across proficiency regions where the examinee population is dense. Let $\hat{\theta}_X$ be the ability estimate corresponding to the observed NC score *X* obtained by the ENC scoring method and $Var\left(\hat{\theta}_X|\theta\right)$ be the squared conditional standard error of the $\hat{\theta}_X$ given a true ability of $\theta$. The marginal test reliability $\bar{\rho}$ is computed as:

$$\bar{\rho} = \frac{\sigma_\theta^2 - \bar{\sigma}_{\theta|\theta}^2}{\sigma_\theta^2}, \qquad (4.2.9)$$

$$\bar{\sigma}_{\theta|\theta}^2 = \sum_{\theta_i} Var(\theta_X \mid \theta_i)\varphi(\theta_i), \qquad (4.2.10)$$

where $\varphi(\theta_i)$ denotes a normalized density of population distribution at $\theta_i$ and $\sigma_\theta^2$ is the variance of population distribution. When a population is assumed to have a standard normal distribution, $\bar{\rho} = 1 - \bar{\sigma}_{\hat{\theta}|\theta}^2$. With this objective function, an optimal MST design should have the maximized marginal test reliability.

A second objective function is the average of CSEEs (across some proficiency region) which prioritizes precise measurement across a broader range of the ability scale. This function tends to promote more precise measurement in the tails of the distribution in comparison to the marginal test reliability objective function. The average of CSEEs across *N* number of $\theta$ points is expressed as:

$$\frac{1}{N}\sum_{\theta_L \le \theta_i \le \theta_U} \sqrt{Var(\hat{\theta}_X \mid \theta_i)}, \qquad (4.2.11)$$

where $\theta_L$ and $\theta_U$ indicate the lower and the upper bounds of the $\theta_i$ given a fixed range. As marginal test reliability, an optimal MST design is found when the average of CSEEs is minimized.

Finally, a maximum CSEE objective function prioritizes designs that avoid imprecisely measuring proficiency value within a given proficiency range. This objective function is given by:

$$\max_{\theta_L \le \theta_i \le \theta_U} \left\{ CSEE(\theta_X \mid \theta_i) \right\}, \tag{4.2.12}$$

where $CSEE\left(\hat{\theta}_X|\theta_i\right)$ is the conditional standard error of $\hat{\theta}_X$ given $\theta_i$. With this objective function, an optimal MST design should have the minimized maximum CSEE.

Since each objective function serves for a different definition of measurement optimality, it is expected that the design characteristic of optimally designed MSTs will vary according to the objective function.

## 4.2 Application to a Real Item Pool

To show that the proposed process is practical and works well, the process described above was applied to a real item pool of a large-scale assessment to find optimal MST designs. Once the optimal designs were found according to each of the three objective functions, the characteristics of the MSTs (e.g., partition of items and other constraints, RIFs, and MIFs) were examined to make a final decision. In the application, it was expected that the optimal design would vary according to the objective function, which is promising because practitioners could flexibly choose the objective function depending on the testing purpose.

### 4.2.1 Design of Application

For the application, two item pools – 200 and 400 items – were used to assume a small and moderate number of items in the MST assembly pools. Also, a fixed number of shapes of MST configurations and test length have been selected to restrict a scope of design space being searched. Among various MST configurations, three widely-used configurations were chosen to represent common practice: the 1-3, 1-2-2, and 1-3-3 MSTs. Two test levels of lengths (32 and 60 items) were used for the short and long test length conditions in MST. Within the restricted design space, the optimal MST design with good overall psychometric outcomes were driven using the three objective functions: (1) maximizing marginal test reliability, (2) minimizing the average of CSEEs given a fixed range of $\theta$, and (3) minimizing the maximum CSEE within a given range of $\theta$.

### 4.2.2 Item Pool

For the two items pools, two hundred and four hundred items calibrated with the IRT 3PL model were randomly selected from an item pool of Massachusetts Adult Proficiency Test – College and Career Readiness (MAPT-CCR) for Mathematics. The MAPT-CCR is a specially designed 5-5-5-5-5-5 MST to measure knowledge and skills in mathematics and reading of adult basic learners in Massachusetts so that their progress in meeting educational goals can be evaluated (Zenisky et al., 2018). Since the test covers a wide range of proficiency levels (i.e., five different difficulty levels in each stage), the item pool contains relatively many items which measures low and high levels of examinee proficiencies compared to other credentialing tests. Table 4 summarizes the descriptive statistics of the item parameters for the two item pools used in the application. The MAPT-CCR for Mathematics measures two dimensions: content strand and cognitive skill dimensions. Instead of using the real item properties, however, four categories of content strand and three categories of cognitive skill were randomly allocated to the items in the pools with the proportion of [.25, .25, .25, .25] and [.25, .25, .50], respectively.

For a test assembly, all MSTs were required to have the same proportions of contend strand and cognitive skill categories as the item pools. In addition, no item overlap across stages and the same test information between two adjacent modules at the RDP were constrained. To prevent an empty module during the ATA, a minimum length of modules was imposed so that each module could have at least 20% of items in the total test.

### 4.2.3 Analysis

All assembled MSTs were scored using the ENC scoring method where the minimum and maximum $\theta$s were set to $-5.0$ and 5.0, respectively. For routing, the DPI rule was employed with the ENC scoring. The CSEEs were computed at the ability points from $-4.0$ to 4.0 in increments of 0.1 using the analytic method. Thus, marginal test reliability for each MST design was calculated from those CSEE values. The average of CSEEs and the maximum of CSEE were obtained from the $\theta$ range of $[-2.0, 2.0]$.

To systematically search for an optimal MST design, the RDPs were varied differently depending on the MST panel configuration. For the 1-3 and 1-3-3 MSTs, the RDP between low- and mid-proficiency levels was intentionally varied within the $\theta$

interval of $[-0.8, -0.1]$ with increments of 0.02 and the RDP between mid- and high-proficiency levels was varied within the $\theta$ interval of [0.1, 0.8] with increments of 0.02. For the 1-2-2 MST, the RDP was allowed to shift within the $\theta$ interval of $[-0.7, 0.7]$ in increments of 0.02.

To use MIP method in the ATA process, a package "lpSolveAPI" (Diao & van der Linden, 2011; Konis, 2009) of R software (R Core Team, 2016) was used. Given each combination of design variables, an MST panel was assembled with a time limit of 3 minutes. The sub-optimal assembly result, which refers to satisfying the objective of the ATA model, that was achieved within the time limit was considered the final assembled MST. All other procedures were conducted with written R code.

### 4.2.4 Results of Application

In each condition of test lengths and item pools, 71 different panels were constructed for the 1-2-2 MST and 1,296 different panels were built for each of the 1-3 and 1-3-3 MST by varying the RDPs. Note that for the 32-item 1-3-3 MSTs, the solver failed to find the solutions of ATA models for few cases of RDPs. Since the number of the failed cases were relatively small (i.e., 9 (0.7%) and 53 (4.1%) cases for the 200- and 400-item pools, respectively), further analyses were carried out without those cases.

#### 4.2.4.1 Examination of Three Objective Functions

##### 4.2.4.1.1 Summary Statistics of Three Objective Functions

Tables 5 through 7 display summary statistics of three objective functions – marginal test reliability, average of CSEEs, and maximum CSEE, respectively –

according to each condition of design variables (i.e., the pool size, test length, and panel configuration). In each condition, the total number of successfully assembled MSTs were presented as well.

Table 5 shows the summary statistics for the first objective function of marginal test reliability. Given the same condition of design variables, three MST panel configurations produced similar summary statistics though the means, maximums, and minimums of the reliabilities were relatively large for the 1-3 and the 1-3-3 MSTs and relatively small for the 1-2-2 MST. Not surprisingly, the tests with more items generally tended to have better reliabilities because the mean, maximum, and minimum statistics for the 60-item test were consistently higher than those for the 32-item test in the two item pools. For example, in the case of the 200-item pool, the means of the reliabilities of the three panel configurations with 60 items ranged from 0.888 to 0.889 while those with 32 items ranged from 0.843 to 0.845. In addition to test length, generally the item pool size affected the marginal test reliability; the larger item pool, the better marginal reliability was, provided that other conditions were the same.

Recall that the definition of measurement optimality will differ by the objective function. When the objective function of marginal test reliability is applied, examining the maximum value among the summary statistics is important to decide the optimal design of an MST. For the 200-item pool, the maximum reliabilities of the three panel configurations were about 0.86 and 0.90 for the 32-item tests and the 60-item tests, respectively and for 400-item pool, those were about 0.88 and 0.92 for the 32-item test and the 60-item test, respectively.

Table 6 presents the summary statistics for the second objective function of average of CSEEs. The means, maximums, and minimums of the average of CSEEs were close to each of the three MST panel configurations when the other design conditions were controlled but, nonetheless, those statistics were relatively small for the 1-3 and the 1-3-3 MSTs and relatively large for the 1-2-2 MST. It can be seen that generally the more items in a test and item pool, the smaller the three summary statistics of the average of CSEEs. For instance, in the case of the 400-item pool, the mean values of the three panel configurations with 60 items were clearly smaller (0.298 to 0.313) than those with 32 items (0.387 to 0.407). When the objective function of the average of CSEEs is used, the optimal design of MST should have the minimized average. Under the 200-item pool, the minimum averages of the three panel configurations were about 0.39 and 0.32 for the 32-item and 60-item tests, respectively and under the 400-item pool, those were about 0.37 and 0.28 for the 32-item and 60-item tests, respectively.

Table 7 shows the summary statistics for the last objective function of maximum CSEE. Unlike the results from the previous two objective functions, it is apparent that the 1-3 and the 1-3-3 MSTs had much lower means and maximums for the maximum CSEE than the 1-2-2 MST, given the same conditions of other design variables. Particularly, when the tests had 32 items under the 400-item pool, the maximum statistics of the maximum CSEE of the 1-3 and 1-3-3 MSTs were about 0.60 whereas those of the 1-2-2 was about 1.13. The three MST panel configurations, however resulted in the minimum statistics of the maximum CSEE close to each other though they were relatively small for the 1-3 and the 1-3-3 MSTs and relatively large for the 1-2-2 MST. Again, as a test and an item pool had more items, it seemed that generally the maximum CSEEs decreased because the means, maximums, and minimums of the 60-item test were consistently smaller than those of the 32-item test in both item pool sizes. Regarding the measurement optimality of MST designs based on the maximum CSEE, it is necessary to select an MST design which has the minimized maximum CSEE. Under the 200-item pool, the minimum of the maximum CSEE of the three panel configurations were about 0.67 and 0.62 for the 32-item and 60-item tests, respectively and under the 400-item pool, those were about 0.60 and 0.47 for the 32-item and 60-item tests, respectively.

#### 4.2.4.1.2 Association Between RDPs and Three Objective Functions

In each condition of the design variables, the MSTs were systematically assembled by varying the RDPs within a specific range of the $\theta$ scale. Accordingly, one may wonder whether there exist any notable relations between the RDPs and the three objective functions. For this reason, the association between the RDPs and each of the three objective functions was examined by means of a scatter plot before selecting an optimal MST design. Figures 11 through 14 display scatter plots between the RDPs and each of the three objective functions according to the three panel configurations in different design conditions of test length and item pool size. In the four figures, each column represents the MST panel configuration and each row indicates the objective function. Recall that one RDP was used for the 1-2-2 MST and two RDPs were used for the 1-3 and 1-3-3 MSTs.

For the 1-3 and 1-3-3 MSTs, weak linear relations were observed between the RDPs and two objective functions of the marginal test reliability and average of CSEEs across all design conditions. More specifically, the marginal test reliability tended to increase slightly and the average of CSEEs was likely to decrease slightly as the two RDPs of both the 1-3 and 1-3-3 MSTs moved from low to high levels in the $\theta$ scale. When it comes to the comparison of two test lengths, the 60-item test showed relatively stronger linear association compared to the 32-item test given the same pool size. No remarkable feature was observed in the relation between the RDPs of the two panel configurations and the maximum CSEE.

For the 1-2-2 MST, it seems that the RDP had nonlinear associations with the three objective functions for all design conditions of test length and item pool size. First, greater values of the marginal test reliability and smaller values of the average of CSEEs were shown as the RDPs were located at around the middle of the $\theta$ scale. This pattern became more noticeable when a test length was 32 (see Table 11 and 13). Second, the nonlinear association was much clearer between the RDPs and maximum CSEE. As the RDP shifted from low to high levels in the $\theta$ scale, the maximum CSEE increased rapidly. As a result, the optimized values were always shown at low levels of the $\theta$ scale regardless of the deign conditions.

However, it should be noted that the results of the association between the RDPs and the three objective functions in this study may not be generalized to other testing circumstances.

#### 4.2.4.2 Decision of Optimal MST Designs

As noted, an optimal MST design would vary depending on which objective function is used to define a measurement optimality. The simplest way of deciding the optimal MST design is to select the one that has the best measurement optimality for each of the objective functions. In other words, the optimal design could be the one that has the highest value when the marginal test reliability was chosen as the objective function and the smallest value when the average of CSEEs or the maximum CSEE was used as the objective function.

However, deciding the best optimal MST design solely based on the objective function value is not recommended because the selected MST design might have inappropriate characteristics of the MIFs or RIFs and unacceptable decomposition results of test-level constraints into modules. For example, an RIF may not provide superior psychometric properties (e.g., test information) for its targeted subpopulations compared to other RIFs which represent other targeted subpopulations. In addition, one may not want modules at a certain stage to have extremely large or small proportions of items at some content categories even though the selected MST design has the optimized value of the objective function. Therefore, further examination is necessary to review several characteristics (e.g., MIFs, RIFs, and the partition of items and other constraints) of assembled MSTs to decide the optimal design of an MST. It is highly recommended to conduct the further examination on a few of the best assembled MSTs for each of the objective functions instead of only one best assembled MST. From this strategy, test developers have more alternatives so that they can choose an MST design by taking into account other characteristics of an MST as well as optimal measurement properties.

Tables 8 through 11 present the MSTs with the top eight optimality values for each of the three objective functions and corresponding RDPs to those MSTs among all assembled MSTs across the three panel configurations. Each table includes the results under different conditions of the design variables (i.e., test length and item pool size). Of course, the number of selected MSTs based on the objective function values could vary by test circumstances and the test developers' intentions. Regardless of the conditions, the top eight MSTs had almost the same measurement optimality values for each objective function. For instance, for the 32-item 1-3-3 MST under the 200-item pool, the differences in the marginal test reliability, average of CSEEs, and maximum CSEEs between the first and the eighth MSTs were only 0.003, 0.003, and 0.005, respectively (see Table 8). Similar results were found in other conditions of panel configurations, test length, and item pool. Hence, it seems to be reasonable to decide an optimal MST design as the one that shows better characteristics of an MST among the top eight MSTs for each objective function through further examination.

Not surprisingly, when the top eight MSTs were selected according to each of the two objective functions of the marginal test reliability and the average of CSEEs, the eight selected MSTs under each of the two objective functions always shared several of the same MSTs which had the same RDPs in each condition of the design variables. Note that since the top-down assembly finds the one best MST design given an RDP (e.g., in the case of the 1-2-2 MST) or a pair of RDPs (e.g., in the cases of the 1-3 and 1-3-3 MSTs), if the assembled MSTs share the same RPDs, this indicates they have exactly the same design (e.g., test and module length, RIFs, and MIFs). For example, in the case of the 32-item test under the 400-item pool, the two objective functions had five of the same tests for the 1-2-2 MST, three of the same tests for the 1-3 MST, and four of the same tests for the 1-3-3 MST (see Table 10). Similar results can be seen in other conditions with respect to test length and item pool. This can be explained by the fact that the nature of the two objective functions are close to each other. Essentially, both the marginal test reliability and the average of CSEEs were associated with the average of the conditional variance of abilities. The difference between them is that the former one involves the computation of the weighted conditional variance of abilities using the population distribution while the latter one requires the unweighted conditional variance of abilities.

Now, the next step is reviewing several characteristics for the selected MSTs to decide the final optimal design of an MST. In this study, the partition of items and content constraints, MIFs, and RIFs were examined for the top four selected MSTs according to each of the three objective functions instead of the top eight selected MSTs for illustrative purposes. As already seen in the previous section, the 60-item test had better objective values for each of the three objective functions under the 400-item pool regardless of MST panel configurations (see Tables 5 through 7). Therefore, the review process was illustrated only for the 60-item test under the 400-item pool. Since the marginal test reliability and average of CSEEs served for similar definitions of measurement optimality, if not identical, the optimal designs were selected based on the objective functions of the marginal test reliability and maximum CSEE.

#### 4.2.4.2.1 Optimal MST Design based on Marginal Test Reliability

Tables 12 through 14 display the partition of items and content constraints for the top four selected 60-item MSTs under the 400-item pool based on the objective function of marginal test reliability. The three tables show the results for the 1-2-2, 1-3, and 1-3-3 MSTs, respectively. Looking at any single MST, each of the test-level specifications (i.e., test length, content strand, and cognitive skill) was uniquely partitioned into modules across stages, satisfying the constraint requirements at the test level when the partitions were summed up. Also, modules in the same stage had identical distributions of items for each test-level constraint, which is consistent with the results in Luo and Kim (2018). Note that the constraint that modules at each stage must contain a number of items equal to at least 20% of total test length was well satisfied for every assembled MSTs.

Concerning the module length of the three panel configurations, the interesting results were found. In the 1-2-2 MST, it seems that items were relatively evenly distributed across three stages for the four selected tests, though three tests contain the largest number of items at the last stage (i.e., the third stage) except the test with the RDP of 0.40 (see Table 12). In the 1-3 and 1-3-3 MSTs, meanwhile, it is prominent that all selected tests showed that the module at the first stage had the longest module length (see Table 13 and 14).

Among all selected designs across the three panel configurations, only the ATA solution of the 1-2-2 MST with the RDP of 0.40 yielded zero item for the fourth category of the content strand in the routing module (see Table 12). If test developers desire that all modules in an MST contains at least a few numbers of items in each content category, any MST with zero items in a certain content category would not be used in operational testing programs. Other than this test, no special problem was observed for all other selected MSTs in terms of the partition of item and content constraints.

Figures 15 through 17 show the RIFs for the top four selected 60-item MSTs under the 400-item pool based on the objective function of marginal test reliability. The three figures present the results for the 1-2-2, 1-3, and 1-3-3 MSTs, respectively. Several interesting features were found from the RIFs of the selected MSTs.

First, all selected MSTs exhibited a clear separation of RIFs of the primary routes. More specifically, two primary RIFs of the four 1-2-2 MSTs and two adjacent primary RIFs of the 1-3 and 1-3-3 MSTs intersected at around their corresponding RDPs and the primary RIFs of each MST which represent different regions of the targeted subpopulations were well distinguished. For example, in the top left panel of Figure 15, the two primary RIFs (i.e., 1M-2E-3E and 1M-2M-3M) of the 1-2-2 MST crossed at around $\theta = -0.46$ and each of them showed better information than the other primary RIF at its defined region of the targeted subpopulation.

Second, when it comes to the three-stage MSTs (i.e., 1-2-2 and 1-3-3 MSTs), the RIFs of the secondary routes were well differentiated from those of the primary routes and the secondary routes which were intended to represent the same region of the targeted subpopulation had similar shapes of the RIFs. In the top left panel of Figure 17, for instance, two secondary RIFs of 1M-2E-3M and 1M-2M-3E were similar to each other and the other two secondary RIFs of 1M-2M-3H and 1M-2H-3M were similar to each other. In addition, the left two secondary RIFs and the right two secondary RIFs were distinctly separated from the three primary RIFs (i.e., 1M-2E-3E, 1M-2M-3M, and 1M-2H -3H). This feature of the clear separation between RIFs is the benefit of the top-down assembly approach because all routes would provide superior measurement precision for their targeted subpopulations in different areas of the $\theta$ scale (Luo & Kim, 2018).

For the 1-2-2 MST, it seems that the secondary RIFs were not as high as the primary RIFs. Due to this fact, one may be concerned about that the precision of ability estimation would be less accurate for examines who take the secondary routes than for those who take the primary routes. As can be seen in all panels of Figure 15, however, the secondary RIFs exhibited the peak information values greater than 15 at around the RDPs, meaning that the CSEEs at those regions are less than 0.26, which is a reasonably good precision of ability estimate because under CAT with the variable length, the standard errors of ability estimates between 0.2 and 0.3 have been commonly used as the prespecified criteria of the stopping rule in many previous research (e.g., Choi, Grady, & Dodd, 2011; Diao & Ren, 2018; Dodd, Kock, & De Ayala, 1993). For the 1-3-3 MST, the RIFs of the secondary routes showed very similar levels of information as those of the first (1M-2E-3E) and third primary (1M-2H-3H) routes at the regions of their targeted subpopulations, implying that measurement precision of the secondary routes are as close as those of the two primary routes at those areas.

Third, an imbalance of information between the primary RIFs was observed for all three MST panel configurations. This feature is best illustrated in the four 1-3 MSTs (see Figure 16) where the primary RIFs (1M-3M) of medium-difficulty level was the highest and the primary RIFs (1M-3E) of hard-difficulty level was the lowest. This could mean that the 400-item pool has abundant information for the medium ability levels of examinees whereas it contains relatively insufficient information for the low and high ability levels of examinees.

With respect to the RIFs in Figures 15 through 17, it seems that all of the selected MSTs have the routes which have the acceptable characteristics of RIFs and well represent their targeted subpopulations in the different regions of the θ scale.

Figures 18 through 20 show the MIFs for the top four selected 60-item MSTs under the 400-item pool based on the objective function of marginal test reliability. The three figures present the results for the 1-2-2, 1-3, and 1-3-3 MSTs, respectively. By reviewing the characteristics of MIFs across stages, the optimized psychometric properties of modules assembled by the top-down approach can be observed.

Similar to the results of the examination on the RIFs, all of the selected MSTs showed that the two adjacent MIFs at each stage were clearly separated, intersecting at the corresponding RDP. This feature allows different difficulty modules at the same stage to provide better measurement precision of ability estimation in their corresponding areas of the targeted subpopulations. In Figure 20, for example, the easy-difficulty modules at the second stage (2E) and third stage (3E) in the 1-3-3 MST had superior information at the left side of the first RDP (e.g., $\theta \leq -0.46$ for MST 1), the medium-difficulty module at the both stages showed higher information than the easy- and high-difficulty level modules between two RDPs (e.g., $-0.46 \leq \theta \leq 0.74$ for MST 1), and the high-difficulty modules at the both stages exhibited the highest information at the right side of the second RDP (e.g., $\theta \geq 0.74$ for MST 1). All other selected MSTs across three panel configurations produced similar results. However, the clear separation of the MIFs obtained from the top-down approach was not observed at the second stage of the 1-2-2 and 1-3-3 MSTs in Luo and Kim (2018); that is, the MIFs of different difficulty modules at the second stage were hardly distinguished. The different results might be attributed to the use of different route mapping strategies as explained in the previous section. Further explanation about the difference is discussed in a later section.

In addition, as shown in Luo and Kim (2018), it was found that MIFs at a certain stage compensated for MIFs at different stages. The four 1-3 MSTs in Figure 19 best illustrated this pattern. In all four MSTs, the MIF at the first stage had relatively low information at higher levels of the $\theta$ scale and relatively high information at the middle and lower levels of the $\theta$ scale. Those features of the MIF at the first stage were compensated for by an opposite trend of the MIFs at the second stage. The compensation of MIFs across stages happened because modules at different stages were considered collectively in the top-down assembly approach (Luo & Kim, 2018).

No specific problem was found for the MIFs in Figures 18 through 20. It seems that all of the selected MSTs consisted of well assembled modules in terms of psychometric properties of MIFs as expected.

In this section, three characteristics of MSTs, which are the partition of items and content constraints, the RIFs, and MIFs, were reviewed for the top four selected MSTs across the three panel configurations under the condition of the 60-item test and the 400-item pool to decide the final optimal design of MSTs based on the objective function of the marginal test reliability. Regarding the partition of content constraints, only one MST (i.e., the 1-2-2 MST with RDP of 0.40) had no item at a certain category of content strand. Other than this case, all other selected MSTs showed that items were well distributed to modules and each content category. In addition, the RIFs and MIFs exhibited superior psychometric properties in their corresponding regions of the targeted subpopulations. According to the examination results, therefore, it is expected that most of the selected MSTs would perform well, regardless of the panel configurations, for the intended testing purpose in this study, which is accurately measuring examinees' proficiencies over a wide range of the $\theta$ scale. If it is necessary to decide the one best MST design in each panel configuration, it would be reasonable to choose the MST that has the highest marginal test reliability.

#### 4.2.4.2.2 Optimal MST Design based on Maximum CSEE

Tables 15 through 17 display the partition of items and content constraints for the top four selected 60-item MSTs under the 400-item pool based on the objective function of maximum CSEE. The three tables show the results for the 1-2-2, 1-3, and 1-3-3 MSTs, respectively. As the partition results of marginal test reliability, each of the selected MSTs satisfied the test-level constraints in the ATA model, having a unique decomposition of items into module and content categories across stages and satisfying a minimum module length of 20% items per each stage.

With respect to module length, however, the selected MSTs based on the maximum CSEE exhibited different patterns of the results from those based on the marginal test reliability. Specifically, the longest module length occurred at the earlier stages (i.e., the first and second stages for the 1-2-2 and 1-3-3 MSTs and the first stage for the 1-3 MSTs) for all selected MST whereas the 1-2-2 MSTs resulted in opposite trend when the top MSTs were selected based on the marginal test reliability.

Among the top four selected MSTs across the three panel configurations, only two 1-2-2 MSTs with RDPs of −0.56 and -0.68, respectively, had no item at certain content categories (see Table 12). For example, for the 1-2-2 MST with the RDP of -0.56, the ATA solution did not allocate any item to the third category of the content strand. Except for the two selected MSTs, no specific problem was found for other selected MSTs in terms of the partition of item and content constraints.

Figures 21 through 23 show the RIFs for the top four selected 60-item MSTs under the 400-item pool based on the objective function of maximum CSEE. The three figures present the results for the 1-2-2, 1-3, and 1-3-3 MSTs, respectively. Although the four selected MSTs in each of the three panel configurations had the smallest maximum CSEEs among all assembled MSTs, it seems that their RIFs which were mapped onto the different regions of the targeted subpopulations were not clearly separated for most of the selected MSTs. Especially, the unclear separations between the RIFs were severe at low levels of the $\theta$ scale regardless of the panel configurations.

For example, for the first 1-2-2 MST with the RDP of $-0.64$ and the fourth 1-2-2 MST with the RDP of $-0.68$, two primary RIFs (i.e., 1M-2E-3E and 1M-2H-3H) were hardly differentiated from the two secondary RIFs (i.e., 1M-2E-3H and 1M-2H-3E) at the $\theta$ scale below the RDPs (see Figure 21). In the case of the 1-3 MST, the first three selected MSTs showed that two primary RIFs (i.e., 1M-2E and 1M-2M) of the easy- and medium-difficulty routes were almost indistinguishable at low levels of the $\theta$ scale (see Figure 22). The problem becomes more serious in the case of the 1-3-3 MST. For all four selected MSTs, it was very hard to distinguish the secondary RIFs of the 1M-2E-3M and 1M-2M-3E from the two primary RIFs of the 1M-2E-3E and 1M-2M-3M at the $\theta$ scale below the first RDPs (see Figure 23). Especially, for the first MST with the RDPs of $-0.76$ and 0.78, the primary route of 1M-2E-3E and the secondary route of 1M-2E-3M had the same RIFs and the primary RIF of 1M-2M-3M and the secondary RIF of 1M-2M-3E also had identical RIFs (see the top left panel of Figure 23).

Figures 24 through 26 show the MIFs for the top four selected 60-item MSTs under the 400-item pool based on the objective function of maximum CSEE. The three figures present the results for the 1-2-2, 1-3, and 1-3-3 MSTs, respectively. Similar to the MIF results of the selected MSTs based on the marginal test reliability, the compensation of the MIFs across stages was observed for all selected MSTs based on the maximum

CSEE. Although most of the selected MSTs exhibited that the different difficulty MIFs could be separated by the RDPs, however, the separation was not so distinct as much as the MIFs of the selected MSTs based on the marginal test reliability. In addition, several MSTs resulted in two adjacent MIFs very close to each other at low levels of the $\theta$ scale, implying that those two modules provide similar levels of information at that region. This could explain the reason that the different RIFs of the selected MSTs based on the maximum CSEE were hardly differentiated at low levels of the $\theta$ scale. For example, the first 1-3-3 MST with the RPDs of $-0.76$ and 0.78 had the identical MIFs of the easy- and medium-difficulty modules (i.e., 3E and 3M) at the third stage (see the first column of Figure 26). From a close examination of the assembled MST, it was found that the two modules at the third stage had exactly the same items. This is the reason that the first 1-3-3 MST resulted in the primary RIF of 1M-2E-3E which was the same with the secondary RIF of 1M-2E-3M and the primary RIF of 1M-2M-3M which was the same with the secondary RIF of 1M-2M-3E (see the top left panel of Figure 23).

In this section, three characteristics of MSTs, which are the partition of items and content constraints, the RIFs, and MIFs, were reviewed for the top four selected MSTs across the three panel configurations under the condition of the 60-item test and the 400-item pool to decide the final optimal design of MSTs based on the objective function of maximum CSEE. When it comes to the partition of items and content constraints, the two selected 1-2-2 MSTs (i.e., the tests with RDPs $-0.56$ and $-0.68$, respectively) contained no item at certain content categories. But, the ATA solutions for all the selected MSTs successfully met the test-level constraints. However, the assembly results in terms of the RIFs were unsatisfactory because most of the selected MSTs had a problem that their

RIFs, which were mapped onto different regions of the targeted subpopulations, were not clearly separated at low levels of the $\theta$ scale. This problem was mainly attributed to the fact that the MIFs of different difficulty modules subsequent to the first stage were not well differentiated at that region.

In this study, it was expected that each route of the optimal MST design should have superior psychometric properties in its corresponding ability region of the targeted subpopulation so that the test is able to precisely estimate examinees' proficiencies over a wide range of the $\theta$ scale. Considering the review of the three characteristics for the top four selected MSTs across the three panel configurations, the use of the maximum CSEE as the objective function might not be appropriate in light of the expectation and testing purpose assumed in this study. Even the best selected MST with the minimized maximum CSEE value in each of the three panel configurations did not seem to work properly for the testing purpose. Therefore, it would be better to find the optimal design of an MST based on other objective functions such as the marginal test reliability and average of CSEEs rather than to decide the optimal design based on the maximum CSEE in this study.

## 4.3 Discussion

Study 2 proposed a process of finding an MST design that has optimal measurement properties given a specific set of testing circumstances. The process consists of four important features: (1) setting a testing circumstance and MST design space, (2) systematically searching the MST design space using a top-down approach, (3) analytically evaluating measurement performance of the MST, and (4) computing objective functions.

The process of discovering an optimally designed MST was applied to a real item pool from a large-scale assessment to show that it would perform well in practice. Given the context of the item pool and the conditions of the design variables under which the application study was conducted, the study revealed the following major findings. First, generally the longer test length and the larger item pool size, the better measurement values for the three objective functions. This trend was more clearly observed for the summary statistic being examined to find the optimal MST design under each objective function. For example, when the marginal test reliability was used as the objective function, examining the maximum statistic of the marginal test reliabilities among all of the assembled MSTs is important to select the optimal MST design. Also, when the average of CSEEs and maximum CSEE were employed as the objective functions, examining the minimum statistics of the two objective function values is important to decide the optimal design. Regarding the maximum statistic of the marginal test reliability, the results of the application study showed that as test length and pool size increased, the maximum of the marginal test reliability increased given the same condition of the design variables. When it comes to the minimum statistics of the average of CSEEs and maximum CSEE, as test length and pool size increased, the minimum of the two objective function values decreased given the other design variables were controlled. This finding replicates the results of the previous research in which a test showed better psychometric properties when it had more items and/or item pool size was larger (e.g., Hambleton & Xing, 2006; Luo & Kim, 2018; Wang et al., 2012; Xing & Hambleton, 2004; Zenisky, 2004).

In addition to the effects of test length and item pool size, another interesting result is that no general pattern was observed with respect to module length across stages. When the objective function of the marginal test reliability was used, more items were allocated to the modules at the later stages (i.e., the second or third stages) than to the routing module for the top four 1-2-2 MSTs while the routing module had the most items for the 1-3 and 1-3-3 MSTs. On the other hand, different trends of module length were shown when the objective function of maximum CSEE was employed. The issue regarding how to distribute items to modules across stages has been studied in many previous MST research (e.g., Kim & Plake, 1993; Patsula, 1999; Zheng et al., 2012). For example, Zheng et al. (2012) argued that it was not clear which allocation strategies produced better measurement properties and the results in this study confirms their claim. Therefore, it might be that the testing context (e.g., item pool, testing purpose, and panel configuration) and other factors (e.g., constraints in the ATA model and the type of objective function) play a large role in the ideal condition of module length across stages.

Second, the three objective functions resulted in different optimal MST designs given the same condition of MST panel configuration, test length, and item pool size. Specifically, when the top eight optimal MSTs were selected according to the three objective functions, the designs of the top eight MSTs differed by the objective function (see Tables 8 through 11). Of course, the marginal test reliability and average of CSEEs tended to share several of the same MSTs among their top eight MSTs due to the similar nature in the definitions of measurement optimality. However, even the same MSTs were rank ordered differently with respect of measurement optimality values between the two objective functions. Moreover, the top eight MSTs based on the maximum CSEE were clearly distinguished from those based on the other two objective functions. This stressed again that the optimal design of an MST would vary depending on which objective function is used.

Third, when focusing on the top eight MSTs under each of the objective functions, they had almost the same measurement optimality values of each objective function. For example, for any conditions of the panel configuration, test length, and item pool, the difference in the objective values between the first and eighth MSTs were less than 0.01 regardless of the objective functions (see Tables 8 through 11). These results may not generalize to other testing contexts. Yet, if it is possible to select at least a few best designs among all of the assembled MSTs that have similar measurement optimality values of an objective function, it would provide us more alternatives for the optimal MST designs. Among the alternatives, one may choose the best design through further review of other characteristics of the tests such as the partition of items and content as well as the RIFs and MIFs. In fact, as stressed earlier, an in-depth review of other characteristics of a few of the best assembled MSTs as well as the optimal measurement properties is an essential part of finding an optimally designed MST. This is because even though an MST design has the best optimality value of an objective function, it could have unacceptable characteristics of the MIFs, RIFs, or the decomposition of test-level constraints into modules.

Fourth, it seems that the objective function of marginal test reliability will perform better in finding an optimally designed MST than the maximum CSEE, provided that the testing purpose is to measure examinees' abilities precisely across a wide range of the $\theta$ scale. In the review process of the top four MSTs selected based on the marginal test reliability, most of the selected MSTs showed satisfactory characteristics of the partition of items and content constraints, and a clear separation between the RIFs and the MIFs which were intended to represent different areas of the targeted subpopulations. Accordingly, it was reasonable to decide an optimal MST design as the one that yielded the highest marginal test reliability among the top four MSTs. Since the objective function of average of the CSEE would function similarly as the marginal test reliability, it is expected that it will show good performance in terms of finding an optimal MST design given the same testing purpose assumed in this study. Under the testing context in this study, however, the maximum CSEE did not seem to work well because, in most conditions of the design variables, the top four MSTs based on the maximum CSEE exhibited the RIFs that mapped onto different regions of the targeted subpopulations were not distinctly separated at low levels of the $\theta$ scale. Therefore, each route would not have superior psychometric properties in its corresponding ability region of the targeted subpopulation compared to other routes representing the different targeted subpopulations, leading to make it difficult to achieve the testing purpose of this study.

However, the objective function of the maximum CSEE could be valuable in other testing context. Suppose that an optimal MST needs to be designed in a credentialing exam to ensure classification accuracy and consistency of pass-fail decision. For this purpose, it is important that the test achieves higher precision for examinees in the region of the passing score. In this context, the use of the maximum CSEE as the objective function would be a good choice since a test that has the minimized maximum CSEE at the passing score would ensure the high quality of psychometric properties of the credentialing exam. In many credentialing testing programs, test developers desire to provide examinees categorized to the failed group with detailed diagnostic feedback as well as make reliable and valid pass-fail decision (Hambleton & Xing, 2006). In this case, the overall objective function could be based on a weighted sum of multiple objective functions. For example, a weighted sum of the maximum CSEE and marginal test reliability could be an objective function to satisfy higher classification accuracy at the pass-fail score as well as good quality of measurement precision at a wide range of ability levels. Of course, how to weight each objective function needs to be further examined.

Instead of the three objective functions used in this study, other types of the objective functions could be introduced depending on a specific context and testing purpose. For example, under a licensure or credentialing testing situation, test developers might prioritize the classification accuracy at cut-scores. In this case, the predicted classification accuracy explained in Chapter 3 could be a good candidate for the objective function because it can be readily derived once the CSEEs on the discrete $\theta$ scale are computed. In fact, it would be interesting to investigate if the objective function of predicted classification accuracy works similarly as the maximum CSEE objective function in the licensure setting. Therefore, it is important to use the objective function in the process of finding an optimally designed MST in accordance with the testing purpose. A proper selection of the objective function that fits testing program’s philosophy will increase the generalizability of the suggested process.

Fifth, the top four MSTs selected based on the marginal test reliability revealed that the different difficulty modules at the same stages had clearly distinguished MIFs which is inconsistent with the results in Luo and Kim (2018). Specifically, their simulation results showed that the different difficulty modules at the second stage in the 1-2-2 and 1-3-3 MSTs assembled through the top-down approach exhibited similar shapes of the MIFs. This inconsistency can be explained by the difference of the route mapping strategies between the two studies. In Luo and Kim (2018), homogenous routes of the three-stage MSTs which have the same difficulty level module at the third stage were mapped onto the same targeted subpopulation. Consequently, this strategy allowed the homogenous routes to have similar shapes of the RIFs. To have similar RIFs, given that the homogenous routes share the same module at the third stage, it is inevitable that the different difficulty modules at the second stage will not have distinguished MIF shapes. If the modules at the second stage were clearly separated, the homogenous route would not have similar RIFs. This did not occur in this study because Luo and Kim's route mapping strategy was modified so that the primary and secondary routes could represent the different areas of the targeted subpopulations. It should be noted, however, that the different characteristics of the observed MIFs between two studies do not mean that one of the strategies is superior to the other. Rather, each of them can be alternatively used depending on how to assign the routes the targeted subpopulations.

# CHAPTER 5

# CONCLUSIONS

The purpose of this dissertation was to propose a process of finding an MST design that has optimal measurement properties, given a specific testing context. To discover the optimal MST designs more efficiently and quickly, an efficient strategy was introduced at each of three phases: constructing MSTs, searching design space of an MST, and evaluating the MST performance. For the first phase, a top-down assembly approach was applied in this study. For the second phase, the parameterized design space of an MST was systematically searched. For the third phase, a new analytical evaluation method of MST was proposed.

This dissertation consisted of two studies. Study 1 introduced the new analytical method to evaluate measurement performance of an MST based on the ENC scoring. Using this new approach, measurement precision (i.e., conditional bias and standard errors) of ability estimates and classification accuracy could be derived analytically. The simulation results in Study 1 indicated that the new analytical method produced more exact measurement properties of an MST than the MC-based simulation method as well as more stable and credible measurement precision and classification accuracies than Park et al.’s (2017) analytical approach. Therefore, it was demonstrated that the new analytical method would be an efficient tool, especially in situation where multiple MST designs need to be compared to find a design that has better measurement performance.

Study 2 proposed a process to find an MST design that has optimal measurement properties applying the three efficient strategies including the new analytical method, given a specific set of testing circumstances. The process consists of four important features: (1) setting a testing circumstance and MST design space, (2) systematically searching the MST design space using a top-down approach, (3) analytically evaluating measurement performance of an MST, and (4) computing objective functions. The process based on the four important features was applied to a real item pool from a large-scale assessment.

The results of the application study in Study 2 serve as evidence of the practical feasibility of the proposed process for finding an optimal MST design for operational testing programs. This is mainly due to the use of the three strategies employed in this study. First, the top-down assembly approach made it relatively easy to optimally partition the test-level design parameters, ensuring the best psychometric properties of an MST given a specific set of test-level design variables. Thus, it could minimize the test developers' subjective decisions for the decomposition of test-level design parameters and prevent the ATA model from returning the suboptimal solution of an MST assembly, which could easily occur if the bottom-up approach was used (Luo & Kim, 2018). Second, by systematically shifting the location of the RDPs in the ATA process using the top-down approach, the parametrized design space of an MST was more efficiently searched. This is because the location of the RPDs determines the targeted subpopulations of the routes in an MST, which, in turn, influences the critical psychometric and nonstatistical characteristics of the optimally designed modules through the top-down assembly. Thus, iteratively varying the RPDs allows the ATA algorithm to automatically consider many combinations of design variables without unnecessary burden of test developers. Third, the new analytical evaluation method made the process of discovering an optimally designed MST feasible. Without the analytical method, the implementation of the process would have not been realized mainly because it would take considerably more time and effort to compute the CSEEs for all of the assembled MSTs if the MC-based simulation method was applied, even with a super powerful computing performance. Due to the new analytical evaluation method, it was possible to assess measurement performance of the thousands of assembled MSTs just in a few minutes.

There are several limitations in this study. First, the procedure of finding an optimal design of an MST and its application were aligned with the purpose of a test to measure achievement or growth for the entire testing population, which means that the assembled MST should have good psychometric properties over a wide range of the proficiency scale. Therefore, the ATA model of the top-down approach was structured so that each route in the assembled MST can produce better measurement precision than the others mapped on the different targeted subpopulations. Also, this is the reason why the objective function of the marginal test reliability behaved well in this study compared to the maximum CSEE. Under a different testing context (e.g., credentialing testing programs), however, the definition of an optimal MST design might differ, which leads to different requirements and implementations of the design process. For example, the objectives of RIFs in the ATA model might be set to allow the test to prioritize the classification accuracies at the cut-scores, and the objective function in the process should be carefully chosen to fit the testing purpose.

Second, the suggested process in this study is able to find only one optimal MST panel design. In real testing programs, it is often required to prepare multiple panels of MSTs for practical reasons (e.g., test security and item pool utilization). Although it is feasible to assemble multiple parallel MST panels theoretically using the top-down approach, this dramatically increases the complexity of the ATA optimization problem. Thus, it may take a significant amount of time to build multiple MST panels for only one combination of the design variables or fail to find an optimal solution of the ATA model even with a high-performing solver. Another practical solution might be using an optimally designed MST, once it is found through the proposed process, as a reference panel (Luo & Kim, 2018). Then, multiple optimal panels can be assembled by replicating the reference panel using a bottom-up assembly approach. More specifically, the MIFs and other statistical and nonstatistical constrains of the reference panel can be used as the targets for the bottom-up assembly. Even though the bottom-up assembly is used, however, the item pool may not support the multiple parallel panels of the optimal MST design depending on the size and quality of the pool. Accordingly, whether the replicated parallel MST panels have similar measurement properties with the optimally designed MST panel under various conditions of the item pool would be an interesting topic for future research.

Third, though the top-down approach is more flexible in the designing process of an MST than the bottom-up approach, it imposes a computer more computational burden to solve the sophisticated MIP problem in the ATA process (Luo & Kim, 2018). If more complicated MIP problem under the top-down approach needs to be addressed in practice, it may require the use of powerful commercial solvers (e.g., CPLES, LINGO, and *Gurobi*) or the control parameters in the solver should be tuned appropriately to handle the convergence problem. In fact, the control parameters in lpSolve were adjusted more loosely in this study than the specific settings that Diao and van der Linden (2011) provided so that more MSTs could be successfully assembled. Therefore, it is necessary to examine the effects of the control parameters to find an optimal MST design in future research.

Fourth, the application study was conducted with only limited conditions of the MST panel design variables and a few sets of constraints in the ATA model, assuming just few examples of potential application to testing programs. In a different testing context, different sets of the design variables need to be considered as well as more requirements for the test specification might be constrained in the ATA model (e.g., including more content areas, specifying more enemy items, and using testlets). Therefore, it is recommended for future studies to consider other factors that might affect the results of optimally designed MSTs. For example, a future study may consider using different characteristics of modules, different structure of MST panels (e.g., 1-2-3 and 1-2-3-4 MSTs), different examinee populations (e.g., negatively skewed populations), or item pools with different features from those used in this study. If a new testing program is about to be established, these contextual findings will be helpful. Notwithstanding that these factors can have significant impact on the results of the optimal design of MSTs, it is expected that the proposed process will perform well in various testing contexts.

Fifth, further examinations for several of the best MSTs, which were selected according to each of the three objective functions, were illustrated based on a few criteria, (i.e., the partition of items and content constraints, RIFs, and MIFs). In addition to the three test characteristics, test developers may want to review other important features of the assembled MSTs. For instance, they may desire an MST to have a better overall module and/or route usage by avoiding an excessive usage rate of certain module and/or routes. Also, reducing the routing error, which occurs when an examinee is not navigated to the intended next module due to measurement error, is an important issue in the implementation of the MST. Therefore, more criteria might need to be considered in further studies to find the optimal MST design in practice.

In conclusion, the results of this study provide evidence that the proposed process with the four features can be generalized to more complex and realistic test circumstances to create optimal designs of MST. Perhaps the most important consideration in generalizing the proposed process is the context of the particular testing program. Context will help test developers to envision ideal statistical and nonstatistical characteristics that an optimal design of an MST should possess and guide specific strategies to be used in the proposed process such as setting objectives and constraints in an ATA model using the top-down approach. Therefore, in future research, some of strategies used in the proposed process need to be modified depending on a specific testing purpose as well as other competing strategies should be developed under various testing context.

**Table 4.** Descriptive Statistics of Item Parameters of Two Item Pools in Study 2

| Parameter | Mean | SD | Min | Max |
|---|---|---|---|---|
| *Pool Size* = 200 | | | | |
| a | 1.26 | 0.39 | 0.35 | 2.70 |
| b | 0.41 | 0.92 | -2.02 | 2.38 |
| c | 0.21 | 0.07 | 0.07 | 0.50 |
| *Pool Size* = 400 | | | | |
| a | 1.25 | 0.40 | 0.21 | 2.83 |
| b | 0.44 | 1.00 | -2.32 | 3.57 |
| c | 0.21 | 0.07 | 0.05 | 0.50 |

**Table 5.** Descriptive Statistics of Marginal Test Reliabilities for the Assembled MSTs

| $n_{item}$ | MST | *N* | Mean | SD | Max | Min |
|---|---|---|---|---|---|---|
| *Pool Size* = 200 | | | | | | |
| 32 | 1-2-2 | 71 | 0.843 | 0.014 | 0.862 | 0.808 |
| | 1-3 | 1,296 | 0.849 | 0.007 | 0.864 | 0.811 |
| | 1-3-3 | 1,287 | 0.845 | 0.010 | 0.864 | 0.801 |
| 60 | 1-2-2 | 71 | 0.888 | 0.008 | 0.899 | 0.864 |
| | 1-3 | 1,296 | 0.889 | 0.006 | 0.902 | 0.868 |
| | 1-3-3 | 1,296 | 0.888 | 0.005 | 0.902 | 0.870 |
| *Pool Size* = 400 | | | | | | |
| 32 | 1-2-2 | 71 | 0.849 | 0.030 | 0.875 | 0.719 |
| | 1-3 | 1,296 | 0.859 | 0.007 | 0.875 | 0.827 |
| | 1-3-3 | 1,243 | 0.857 | 0.010 | 0.875 | 0.793 |
| 60 | 1-2-2 | 71 | 0.904 | 0.009 | 0.918 | 0.863 |
| | 1-3 | 1,296 | 0.912 | 0.003 | 0.919 | 0.898 |
| | 1-3-3 | 1,296 | 0.910 | 0.004 | 0.920 | 0.891 |

*Note*. $n_{item}$ = test length; *N* = total number of successfully assembled MSTs.

**Table 6.** Descriptive Statistics of Average of CSEEs for the Assembled MSTs

| $n_{item}$ | MST | *N* | Mean | SD | Max | Min |
|---|---|---|---|---|---|---|
| *Pool Size* = 200 | | | | | | |
| 32 | 1-2-2 | 71 | 0.424 | 0.019 | 0.467 | 0.395 |
| | 1-3 | 1,296 | 0.411 | 0.010 | 0.462 | 0.392 |
| | 1-3-3 | 1,287 | 0.417 | 0.013 | 0.475 | 0.391 |
| 60 | 1-2-2 | 71 | 0.346 | 0.014 | 0.386 | 0.327 |
| | 1-3 | 1,296 | 0.343 | 0.010 | 0.377 | 0.320 |
| | 1-3-3 | 1,296 | 0.343 | 0.009 | 0.374 | 0.319 |
| *Pool Size* = 400 | | | | | | |
| 32 | 1-2-2 | 71 | 0.407 | 0.032 | 0.543 | 0.375 |
| | 1-3 | 1,296 | 0.387 | 0.009 | 0.442 | 0.369 |
| | 1-3-3 | 1,243 | 0.392 | 0.014 | 0.479 | 0.365 |
| 60 | 1-2-2 | 71 | 0.313 | 0.014 | 0.360 | 0.292 |
| | 1-3 | 1,296 | 0.298 | 0.007 | 0.321 | 0.283 |
| | 1-3-3 | 1,296 | 0.300 | 0.008 | 0.332 | 0.284 |

*Note*. $n_{item}$ = test length; *N* = total number of successfully assembled MSTs.

**Table 7.** Descriptive Statistics of Maximum CSEE for the Assembled MSTs

| $n_{item}$ | MST | *N* | Mean | SD | Max | Min |
|---|---|---|---|---|---|---|
| *Pool Size* = 200 | | | | | | |
| 32 | 1-2-2 | 71 | 0.748 | 0.088 | 0.995 | 0.671 |
| | 1-3 | 1,296 | 0.683 | 0.010 | 0.752 | 0.669 |
| | 1-3-3 | 1,287 | 0.686 | 0.015 | 0.868 | 0.662 |
| 60 | 1-2-2 | 71 | 0.676 | 0.060 | 0.936 | 0.633 |
| | 1-3 | 1,296 | 0.638 | 0.002 | 0.664 | 0.624 |
| | 1-3-3 | 1,296 | 0.638 | 0.004 | 0.696 | 0.613 |
| *Pool Size* = 400 | | | | | | |
| 32 | 1-2-2 | 71 | 0.786 | 0.218 | 1.477 | 0.622 |
| | 1-3 | 1,296 | 0.649 | 0.019 | 0.791 | 0.597 |
| | 1-3-3 | 1,243 | 0.657 | 0.036 | 1.059 | 0.593 |
| 60 | 1-2-2 | 71 | 0.642 | 0.140 | 1.126 | 0.473 |
| | 1-3 | 1,296 | 0.505 | 0.029 | 0.602 | 0.473 |
| | 1-3-3 | 1,296 | 0.510 | 0.033 | 0.665 | 0.462 |

*Note*. $n_{item}$ = test length; *N* = total number of successfully assembled MSTs.

**Table 8.** Top Eight MSTs According to Three Objective Functions: 32-Item Test under 200-Item Pool

| | Marginal Test Reliability | | Average of CSEEs | | Maximum CSEE | |
|---|---|---|---|---|---|---|
| MST | Value | RDP | Value | RDP | Value | RDP |
| 1-2-2 | 0.862 | -0.08 | 0.395 | -0.08 | 0.671 | -0.42 |
| | 0.860 | 0.16 | 0.398 | 0.16 | 0.671 | -0.34 |
| | 0.860 | -0.24 | 0.400 | -0.24 | 0.671 | -0.44 |
| | 0.860 | 0.14 | 0.402 | -0.12 | 0.671 | -0.38 |
| | 0.859 | -0.12 | 0.402 | 0.14 | 0.672 | -0.40 |
| | 0.858 | 0.20 | 0.403 | -0.28 | 0.678 | -0.60 |
| | 0.857 | -0.26 | 0.403 | -0.26 | 0.678 | -0.52 |
| | 0.857 | 0.26 | 0.404 | -0.30 | 0.679 | -0.58 |
| 1-3 | 0.864 | -0.30, 0.64 | 0.392 | -0.30, 0.64 | 0.669 | -0.72, 0.26 |
| | 0.862 | -0.16, 0.62 | 0.393 | -0.34, 0.76 | 0.671 | -0.36, 0.18 |
| | 0.862 | -0.12, 0.70 | 0.393 | -0.40, 0.70 | 0.671 | -0.44, 0.12 |
| | 0.861 | -0.10, 0.66 | 0.393 | -0.16, 0.62 | 0.671 | -0.34, 0.44 |
| | 0.861 | -0.14, 0.66 | 0.393 | -0.26, 0.36 | 0.671 | -0.34, 0.26 |
| | 0.861 | -0.18, 0.76 | 0.394 | -0.10, 0.66 | 0.671 | -0.38, 0.16 |
| | 0.861 | -0.40, 0.70 | 0.394 | -0.40, 0.64 | 0.671 | -0.34, 0.14 |
| | 0.861 | -0.32, 0.76 | 0.394 | -0.12, 0.78 | 0.671 | -0.44, 0.10 |
| 1-3-3 | 0.864 | -0.28, 0.72 | 0.391 | -0.28, 0.72 | 0.662 | -0.58, 0.58 |
| | 0.862 | -0.10, 0.38 | 0.392 | -0.34, 0.78 | 0.666 | -0.44, 0.36 |
| | 0.862 | -0.34, 0.78 | 0.392 | -0.14, 0.80 | 0.666 | -0.34, 0.58 |
| | 0.861 | -0.10, 0.12 | 0.393 | -0.30, 0.80 | 0.666 | -0.44, 0.12 |
| | 0.861 | -0.28, 0.18 | 0.393 | -0.26, 0.76 | 0.666 | -0.34, 0.48 |
| | 0.861 | -0.14, 0.80 | 0.394 | -0.46, 0.70 | 0.666 | -0.72, 0.26 |
| | 0.861 | -0.26, 0.30 | 0.394 | -0.14, 0.76 | 0.667 | -0.34, 0.16 |
| | 0.861 | -0.26, 0.76 | 0.394 | -0.10, 0.38 | 0.667 | -0.68, 0.78 |

*Note*. RPD = routing decision point.

**Table 9.** Top Eight MSTs According to Three Objective Functions: 60-Item Test under 200-Item Pool

| | Marginal Test Reliability | | Average of CSEEs | | Maximum CSEE | |
|---|---|---|---|---|---|---|
| MST | Value | RDP | Value | RDP | Value | RDP |
| 1-2-2 | 0.899 | 0.36 | 0.327 | 0.42 | 0.633 | -0.14 |
| | 0.898 | 0.42 | 0.327 | 0.36 | 0.633 | -0.08 |
| | 0.898 | 0.46 | 0.328 | 0.40 | 0.633 | -0.10 |
| | 0.898 | 0.44 | 0.329 | 0.44 | 0.633 | -0.04 |
| | 0.898 | 0.40 | 0.329 | 0.48 | 0.633 | -0.12 |
| | 0.898 | 0.38 | 0.330 | 0.46 | 0.633 | -0.06 |
| | 0.897 | 0.28 | 0.330 | 0.38 | 0.633 | -0.16 |
| | 0.897 | 0.48 | 0.330 | 0.28 | 0.633 | -0.02 |
| 1-3 | 0.902 | -0.14, 0.76 | 0.320 | -0.14, 0.76 | 0.624 | -0.76, 0.76 |
| | 0.901 | -0.16, 0.72 | 0.321 | -0.10, 0.78 | 0.632 | -0.78, 0.34 |
| | 0.901 | -0.10, 0.78 | 0.322 | -0.12, 0.76 | 0.633 | -0.16, 0.16 |
| | 0.901 | -0.12, 0.80 | 0.322 | -0.14, 0.80 | 0.633 | -0.10, 0.60 |
| | 0.901 | -0.14, 0.74 | 0.322 | -0.14, 0.74 | 0.633 | -0.16, 0.24 |
| | 0.901 | -0.12, 0.76 | 0.322 | -0.14, 0.78 | 0.633 | -0.12, 0.10 |
| | 0.900 | -0.14, 0.80 | 0.322 | -0.16, 0.72 | 0.633 | -0.10, 0.38 |
| | 0.900 | -0.14, 0.78 | 0.322 | -0.12, 0.80 | 0.633 | -0.10, 0.34 |
| 1-3-3 | 0.902 | -0.14, 0.68 | 0.319 | -0.14, 0.68 | 0.613 | -0.74, 0.64 |
| | 0.900 | -0.12, 0.66 | 0.321 | -0.12, 0.66 | 0.613 | -0.70, 0.20 |
| | 0.900 | -0.12, 0.78 | 0.322 | -0.12, 0.78 | 0.617 | -0.56, 0.50 |
| | 0.900 | -0.16, 0.68 | 0.323 | -0.16, 0.68 | 0.620 | -0.80, 0.42 |
| | 0.899 | -0.10, 0.18 | 0.324 | -0.34, 0.74 | 0.622 | -0.76, 0.32 |
| | 0.899 | -0.16, 0.72 | 0.324 | -0.16, 0.72 | 0.622 | -0.18, 0.48 |
| | 0.898 | -0.14, 0.66 | 0.324 | -0.18, 0.72 | 0.622 | -0.78, 0.72 |
| | 0.898 | -0.14, 0.76 | 0.325 | -0.28, 0.76 | 0.624 | -0.78, 0.36 |

*Note*. RPD = routing decision point.

**Table 10.** Top Eight MSTs According to Three Objective Functions: 32-Item Test under 400-Item Pool

| | Marginal Test Reliability | | Average of CSEEs | | Maximum CSEE | |
|---|---|---|---|---|---|---|
| MST | Value | RDP | Value | RDP | Value | RDP |
| 1-2-2 | 0.875 | 0.12 | 0.375 | 0.12 | 0.622 | -0.68 |
| | 0.874 | 0.32 | 0.376 | 0.32 | 0.622 | -0.70 |
| | 0.871 | -0.12 | 0.376 | -0.06 | 0.633 | -0.60 |
| | 0.871 | -0.10 | 0.378 | -0.12 | 0.645 | -0.06 |
| | 0.871 | -0.06 | 0.379 | 0.04 | 0.647 | -0.62 |
| | 0.870 | 0.40 | 0.38 | -0.26 | 0.647 | -0.58 |
| | 0.870 | 0.02 | 0.38 | -0.10 | 0.648 | -0.52 |
| | 0.869 | -0.04 | 0.38 | -0.36 | 0.648 | -0.54 |
| 1-3 | 0.875 | -0.14, 0.12 | 0.369 | -0.14, 0.12 | 0.597 | -0.80, 0.70 |
| | 0.872 | -0.10, 0.38 | 0.371 | -0.18, 0.70 | 0.612 | -0.80, 0.14 |
| | 0.872 | -0.44, 0.80 | 0.371 | -0.44, 0.80 | 0.612 | -0.72, 0.20 |
| | 0.872 | -0.12, 0.28 | 0.372 | -0.24, 0.70 | 0.612 | -0.74, 0.38 |
| | 0.872 | -0.42, 0.62 | 0.372 | -0.12, 0.78 | 0.612 | -0.76, 0.22 |
| | 0.872 | -0.12, 0.46 | 0.372 | -0.10, 0.80 | 0.612 | -0.74, 0.12 |
| | 0.871 | -0.16, 0.52 | 0.372 | -0.14, 0.34 | 0.612 | -0.76, 0.24 |
| | 0.871 | -0.40, 0.62 | 0.373 | -0.10, 0.38 | 0.612 | -0.72, 0.12 |
| 1-3-3 | 0.875 | -0.14, 0.72 | 0.365 | -0.46, 0.60 | 0.593 | -0.76, 0.66 |
| | 0.875 | -0.40, 0.78 | 0.367 | -0.14, 0.72 | 0.595 | -0.74, 0.26 |
| | 0.875 | -0.56, 0.70 | 0.369 | -0.58, 0.74 | 0.599 | -0.76, 0.44 |
| | 0.874 | -0.58, 0.74 | 0.369 | -0.50, 0.78 | 0.601 | -0.78, 0.44 |
| | 0.874 | -0.16, 0.56 | 0.369 | -0.40, 0.78 | 0.601 | -0.80, 0.10 |
| | 0.874 | -0.76, 0.54 | 0.369 | -0.56, 0.70 | 0.601 | -0.78, 0.70 |
| | 0.874 | -0.46, 0.60 | 0.369 | -0.42, 0.50 | 0.602 | -0.70, 0.70 |
| | 0.873 | -0.10, 0.50 | 0.37 | -0.60, 0.66 | 0.603 | -0.78, 0.78 |

*Note*. RPD = routing decision point.

**Table 11.** Top Eight MSTs According to Three Objective Functions: 60-Item Test under 400-Item Pool

| | Marginal Test Reliability | | Average of CSEEs | | Maximum CSEE | |
|---|---|---|---|---|---|---|
| MST | Value | RDP | Value | RDP | Value | RDP |
| 1-2-2 | 0.918 | 0.38 | 0.292 | 0.00 | 0.473 | -0.64 |
| | 0.917 | 0.40 | 0.294 | 0.38 | 0.477 | -0.56 |
| | 0.915 | 0.32 | 0.295 | 0.22 | 0.497 | -0.52 |
| | 0.915 | 0.30 | 0.296 | 0.40 | 0.502 | -0.68 |
| | 0.915 | 0.36 | 0.297 | 0.12 | 0.505 | -0.42 |
| | 0.913 | -0.14 | 0.297 | 0.30 | 0.509 | -0.54 |
| | 0.913 | 0.34 | 0.297 | -0.06 | 0.520 | -0.08 |
| | 0.913 | 0.00 | 0.297 | 0.32 | 0.527 | -0.50 |
| 1-3 | 0.919 | -0.46, 0.54 | 0.283 | -0.24, 0.80 | 0.473 | -0.78, 0.40 |
| | 0.919 | -0.52, 0.80 | 0.284 | -0.16, 0.80 | 0.473 | -0.74, 0.10 |
| | 0.919 | -0.54, 0.76 | 0.284 | -0.24, 0.74 | 0.473 | -0.78, 0.20 |
| | 0.919 | -0.58, 0.78 | 0.285 | -0.14, 0.76 | 0.473 | -0.72, 0.68 |
| | 0.919 | -0.42, 0.60 | 0.285 | -0.24, 0.64 | 0.473 | -0.74, 0.44 |
| | 0.919 | -0.42, 0.62 | 0.285 | -0.20, 0.76 | 0.473 | -0.72, 0.42 |
| | 0.918 | -0.42, 0.70 | 0.286 | -0.16, 0.76 | 0.473 | -0.72, 0.38 |
| | 0.918 | -0.42, 0.76 | 0.286 | -0.52, 0.80 | 0.473 | -0.76, 0.12 |
| 1-3-3 | 0.920 | -0.46, 0.74 | 0.284 | -0.14, 0.68 | 0.462 | -0.76, 0.78 |
| | 0.919 | -0.44, 0.72 | 0.285 | -0.18, 0.78 | 0.466 | -0.78, 0.30 |
| | 0.919 | -0.62, 0.78 | 0.285 | -0.18, 0.76 | 0.468 | -0.80, 0.66 |
| | 0.919 | -0.44, 0.68 | 0.286 | -0.14, 0.76 | 0.468 | -0.70, 0.64 |
| | 0.919 | -0.60, 0.74 | 0.286 | -0.18, 0.62 | 0.469 | -0.46, 0.54 |
| | 0.919 | -0.58, 0.64 | 0.287 | -0.24, 0.60 | 0.469 | -0.46, 0.26 |
| | 0.918 | -0.50, 0.74 | 0.287 | -0.12, 0.58 | 0.470 | -0.72, 0.50 |
| | 0.918 | -0.48, 0.78 | 0.287 | -0.10, 0.70 | 0.471 | -0.48, 0.76 |

*Note*. RPD = routing decision point.

**Table 12.** Partition of Items and Constraints of the Top Four MSTs based on Marginal Test Reliability: 1-2-2 MST with 60 Items under 400-Item Pool

| | | | Content Strand | | | | Cognitive | | |
|---|---|---|---|---|---|---|---|---|---|
| RDP | Module | $n_{item}$ | 1 | 2 | 3 | 4 | 1 | 2 | 3 |
| 0.38 | 1M | 21 | 3 | 7 | 4 | 7 | 7 | 7 | 7 |
| | 2E | 17 | 6 | 5 | 4 | 2 | 5 | 3 | 9 |
| | 2H | 17 | 6 | 5 | 4 | 2 | 5 | 3 | 9 |
| | 3E | 22 | 6 | 3 | 7 | 6 | 3 | 5 | 14 |
| | 3H | 22 | 6 | 3 | 7 | 6 | 3 | 5 | 14 |
| 0.40 | 1M | 13 | 3 | 5 | 5 | 0 | 6 | 2 | 5 |
| | 2E | 25 | 5 | 6 | 6 | 8 | 4 | 8 | 13 |
| | 2H | 25 | 5 | 6 | 6 | 8 | 4 | 8 | 13 |
| | 3E | 22 | 7 | 4 | 4 | 7 | 5 | 5 | 12 |
| | 3H | 22 | 7 | 4 | 4 | 7 | 5 | 5 | 12 |
| 0.32 | 1M | 20 | 6 | 6 | 3 | 5 | 4 | 4 | 12 |
| | 2E | 19 | 6 | 4 | 4 | 5 | 4 | 5 | 10 |
| | 2H | 19 | 6 | 4 | 4 | 5 | 4 | 5 | 10 |
| | 3E | 21 | 3 | 5 | 8 | 5 | 7 | 6 | 8 |
| | 3H | 21 | 3 | 5 | 8 | 5 | 7 | 6 | 8 |
| 0.30 | 1M | 16 | 4 | 7 | 2 | 3 | 3 | 5 | 8 |
| | 2E | 21 | 4 | 5 | 7 | 5 | 6 | 5 | 10 |
| | 2H | 21 | 4 | 5 | 7 | 5 | 6 | 5 | 10 |
| | 3E | 23 | 7 | 3 | 6 | 7 | 6 | 5 | 12 |
| | 3H | 23 | 7 | 3 | 6 | 7 | 6 | 5 | 12 |

*Note*. RPD = routing decision point.

**Table 13.** Partition of Items and Constraints of the Top Four MSTs based on Marginal Test Reliability: 1-3 MST with 60 Items under 400-Item Pool

| | | | Content Strand | | | | Cognitive | | |
|---|---|---|---|---|---|---|---|---|---|
| RDP | Module | $n_{item}$ | 1 | 2 | 3 | 4 | 1 | 2 | 3 |
| -0.46, 0.54 | 1M | 39 | 10 | 13 | 6 | 10 | 11 | 11 | 17 |
| | 2E | 21 | 5 | 2 | 9 | 5 | 4 | 4 | 13 |
| | 2M | 21 | 5 | 2 | 9 | 5 | 4 | 4 | 13 |
| | 2H | 21 | 5 | 2 | 9 | 5 | 4 | 4 | 13 |
| -0.52, 0.80 | 1M | 33 | 7 | 8 | 12 | 6 | 8 | 8 | 17 |
| | 2E | 27 | 8 | 7 | 3 | 9 | 7 | 7 | 13 |
| | 2M | 27 | 8 | 7 | 3 | 9 | 7 | 7 | 13 |
| | 2H | 27 | 8 | 7 | 3 | 9 | 7 | 7 | 13 |
| -0.54, 0.76 | 1M | 34 | 10 | 9 | 7 | 8 | 11 | 9 | 14 |
| | 2E | 26 | 5 | 6 | 8 | 7 | 4 | 6 | 16 |
| | 2M | 26 | 5 | 6 | 8 | 7 | 4 | 6 | 16 |
| | 2H | 26 | 5 | 6 | 8 | 7 | 4 | 6 | 16 |
| -0.58, 0.78 | 1M | 37 | 5 | 10 | 10 | 12 | 9 | 10 | 18 |
| | 2E | 23 | 10 | 5 | 5 | 3 | 6 | 5 | 12 |
| | 2M | 23 | 10 | 5 | 5 | 3 | 6 | 5 | 12 |
| | 2H | 23 | 10 | 5 | 5 | 3 | 6 | 5 | 12 |

*Note*. RPD = routing decision point.

**Table 14.** Partition of Items and Constraints of the Top Four MSTs based on Marginal Test Reliability: 1-3-3 MST with 60 Items under 400-Item Pool

| | | | Content Strand | | | | Cognitive | | |
|---|---|---|---|---|---|---|---|---|---|
| RDP | Module | $n_{item}$ | 1 | 2 | 3 | 4 | 1 | 2 | 3 |
| -0.46, 0.74 | 1M | 30 | 7 | 7 | 9 | 7 | 7 | 9 | 14 |
| | 2E | 15 | 3 | 5 | 3 | 4 | 3 | 3 | 9 |
| | 2M | 15 | 3 | 5 | 3 | 4 | 3 | 3 | 9 |
| | 2H | 15 | 3 | 5 | 3 | 4 | 3 | 3 | 9 |
| | 3E | 15 | 5 | 3 | 3 | 4 | 5 | 3 | 7 |
| | 3M | 15 | 5 | 3 | 3 | 4 | 5 | 3 | 7 |
| | 3H | 15 | 5 | 3 | 3 | 4 | 5 | 3 | 7 |
| -0.44, 0.72 | 1M | 27 | 7 | 5 | 8 | 7 | 7 | 6 | 14 |
| | 2E | 17 | 5 | 3 | 3 | 6 | 4 | 5 | 8 |
| | 2M | 17 | 5 | 3 | 3 | 6 | 4 | 5 | 8 |
| | 2H | 17 | 5 | 3 | 3 | 6 | 4 | 5 | 8 |
| | 3E | 16 | 3 | 7 | 4 | 2 | 4 | 4 | 8 |
| | 3M | 16 | 3 | 7 | 4 | 2 | 4 | 4 | 8 |
| | 3H | 16 | 3 | 7 | 4 | 2 | 4 | 4 | 8 |
| -0.62, 0.78 | 1M | 33 | 8 | 10 | 8 | 7 | 8 | 9 | 16 |
| | 2E | 15 | 3 | 3 | 5 | 4 | 4 | 4 | 7 |
| | 2M | 15 | 3 | 3 | 5 | 4 | 4 | 4 | 7 |
| | 2H | 15 | 3 | 3 | 5 | 4 | 4 | 4 | 7 |
| | 3E | 12 | 4 | 2 | 2 | 4 | 3 | 2 | 7 |
| | 3M | 12 | 4 | 2 | 2 | 4 | 3 | 2 | 7 |
| | 3H | 12 | 4 | 2 | 2 | 4 | 3 | 2 | 7 |
| -0.44, 0.68 | 1M | 27 | 7 | 8 | 4 | 8 | 6 | 7 | 14 |
| | 2E | 17 | 4 | 4 | 5 | 4 | 4 | 5 | 8 |
| | 2M | 17 | 4 | 4 | 5 | 4 | 4 | 5 | 8 |
| | 2H | 17 | 4 | 4 | 5 | 4 | 4 | 5 | 8 |
| | 3E | 16 | 4 | 3 | 6 | 3 | 5 | 3 | 8 |
| | 3M | 16 | 4 | 3 | 6 | 3 | 5 | 3 | 8 |
| | 3H | 16 | 4 | 3 | 6 | 3 | 5 | 3 | 8 |

*Note*. RPD = routing decision point.

**Table 15.** Partition of Items and Constraints of the Top Four MSTs based on Maximum CSEE: 1-2-2 MST with 60 Items under 400-Item Pool

| | | | Content Strand | | | | Cognitive | | |
|---|---|---|---|---|---|---|---|---|---|
| RDP | Module | $n_{item}$ | 1 | 2 | 3 | 4 | 1 | 2 | 3 |
| -0.64 | 1M | 25 | 7 | 7 | 4 | 7 | 7 | 7 | 11 |
| | 2E | 20 | 6 | 5 | 6 | 3 | 5 | 2 | 13 |
| | 2H | 20 | 6 | 5 | 6 | 3 | 5 | 2 | 13 |
| | 3E | 15 | 2 | 3 | 5 | 5 | 3 | 6 | 6 |
| | 3H | 15 | 2 | 3 | 5 | 5 | 3 | 6 | 6 |
| -0.56 | 1M | 25 | 5 | 9 | 7 | 4 | 8 | 6 | 11 |
| | 2E | 20 | 4 | 3 | 5 | 8 | 4 | 9 | 7 |
| | 2H | 20 | 4 | 3 | 5 | 8 | 4 | 9 | 7 |
| | 3E | 15 | 6 | 3 | 3 | 3 | 3 | 0 | 12 |
| | 3H | 15 | 6 | 3 | 3 | 3 | 3 | 0 | 12 |
| -0.52 | 1M | 22 | 5 | 4 | 8 | 5 | 6 | 6 | 10 |
| | 2E | 24 | 8 | 6 | 4 | 6 | 6 | 6 | 12 |
| | 2H | 24 | 8 | 6 | 4 | 6 | 6 | 6 | 12 |
| | 3E | 14 | 2 | 5 | 3 | 4 | 3 | 3 | 8 |
| | 3H | 14 | 2 | 5 | 3 | 4 | 3 | 3 | 8 |
| -0.68 | 1M | 36 | 7 | 9 | 11 | 9 | 9 | 9 | 18 |
| | 2E | 12 | 5 | 2 | 4 | 1 | 4 | 2 | 6 |
| | 2H | 12 | 5 | 2 | 4 | 1 | 4 | 2 | 6 |
| | 3E | 12 | 3 | 4 | 0 | 5 | 2 | 4 | 6 |
| | 3H | 12 | 3 | 4 | 0 | 5 | 2 | 4 | 6 |

*Note*. RPD = routing decision point.

**Table 16.** Partition of Items and Constraints of the Top Four MSTs based on Maximum CSEE: 1-3 MST with 60 Items under 400-Item Pool

| | | | Content Strand | | | | Cognitive | | |
|---|---|---|---|---|---|---|---|---|---|
| RDP | Module | $n_{item}$ | 1 | 2 | 3 | 4 | 1 | 2 | 3 |
| -0.78, 0.4 | 1M | 45 | 13 | 12 | 9 | 11 | 14 | 9 | 22 |
| | 2E | 15 | 2 | 3 | 6 | 4 | 1 | 6 | 8 |
| | 2M | 15 | 2 | 3 | 6 | 4 | 1 | 6 | 8 |
| | 2H | 15 | 2 | 3 | 6 | 4 | 1 | 6 | 8 |
| -0.74, 0.1 | 1M | 45 | 11 | 11 | 12 | 11 | 10 | 11 | 24 |
| | 2E | 15 | 4 | 4 | 3 | 4 | 5 | 4 | 6 |
| | 2M | 15 | 4 | 4 | 3 | 4 | 5 | 4 | 6 |
| | 2H | 15 | 4 | 4 | 3 | 4 | 5 | 4 | 6 |
| -0.78, 0.2 | 1M | 46 | 11 | 12 | 12 | 11 | 12 | 8 | 26 |
| | 2E | 14 | 4 | 3 | 3 | 4 | 3 | 7 | 4 |
| | 2M | 14 | 4 | 3 | 3 | 4 | 3 | 7 | 4 |
| | 2H | 14 | 4 | 3 | 3 | 4 | 3 | 7 | 4 |
| -0.72, 0.68 | 1M | 41 | 9 | 10 | 12 | 10 | 10 | 11 | 20 |
| | 2E | 19 | 6 | 5 | 3 | 5 | 5 | 4 | 10 |
| | 2M | 19 | 6 | 5 | 3 | 5 | 5 | 4 | 10 |
| | 2H | 19 | 6 | 5 | 3 | 5 | 5 | 4 | 10 |

*Note*. RPD = routing decision point.

**Table 17.** Partition of Items and Constraints of the Top Four MSTs based on Maximum CSEE: 1-3-3 MST with 60 Items under 400-Item Pool

| | | | Content Strand | | | | Cognitive | | |
|---|---|---|---|---|---|---|---|---|---|
| RDP | Module | $n_{item}$ | 1 | 2 | 3 | 4 | 1 | 2 | 3 |
| -0.76, 0.78 | 1M | 16 | 3 | 7 | 4 | 2 | 5 | 4 | 7 |
| | 2E | 31 | 7 | 6 | 6 | 12 | 8 | 10 | 13 |
| | 2M | 31 | 7 | 6 | 6 | 12 | 8 | 10 | 13 |
| | 2H | 31 | 7 | 6 | 6 | 12 | 8 | 10 | 13 |
| | 3E | 13 | 5 | 2 | 5 | 1 | 2 | 1 | 10 |
| | 3M | 13 | 5 | 2 | 5 | 1 | 2 | 1 | 10 |
| | 3H | 13 | 5 | 2 | 5 | 1 | 2 | 1 | 10 |
| -0.78, 0.3 | 1M | 21 | 8 | 6 | 5 | 2 | 5 | 5 | 11 |
| | 2E | 26 | 5 | 7 | 6 | 8 | 6 | 7 | 13 |
| | 2M | 26 | 5 | 7 | 6 | 8 | 6 | 7 | 13 |
| | 2H | 26 | 5 | 7 | 6 | 8 | 6 | 7 | 13 |
| | 3E | 13 | 2 | 2 | 4 | 5 | 4 | 3 | 6 |
| | 3M | 13 | 2 | 2 | 4 | 5 | 4 | 3 | 6 |
| | 3H | 13 | 2 | 2 | 4 | 5 | 4 | 3 | 6 |
| -0.8, 0.66 | 1M | 31 | 6 | 7 | 10 | 8 | 8 | 8 | 15 |
| | 2E | 14 | 5 | 4 | 2 | 3 | 5 | 3 | 6 |
| | 2M | 14 | 5 | 4 | 2 | 3 | 5 | 3 | 6 |
| | 2H | 14 | 5 | 4 | 2 | 3 | 5 | 3 | 6 |
| | 3E | 15 | 4 | 4 | 3 | 4 | 2 | 4 | 9 |
| | 3M | 15 | 4 | 4 | 3 | 4 | 2 | 4 | 9 |
| | 3H | 15 | 4 | 4 | 3 | 4 | 2 | 4 | 9 |
| -0.7, 0.64 | 1M | 31 | 8 | 9 | 9 | 5 | 9 | 8 | 14 |
| | 2E | 17 | 3 | 4 | 2 | 8 | 4 | 4 | 9 |
| | 2M | 17 | 3 | 4 | 2 | 8 | 4 | 4 | 9 |
| | 2H | 17 | 3 | 4 | 2 | 8 | 4 | 4 | 9 |
| | 3E | 12 | 4 | 2 | 4 | 2 | 2 | 3 | 7 |
| | 3M | 12 | 4 | 2 | 4 | 2 | 2 | 3 | 7 |
| | 3H | 12 | 4 | 2 | 4 | 2 | 2 | 3 | 7 |

*Note*. RPD = routing decision point.

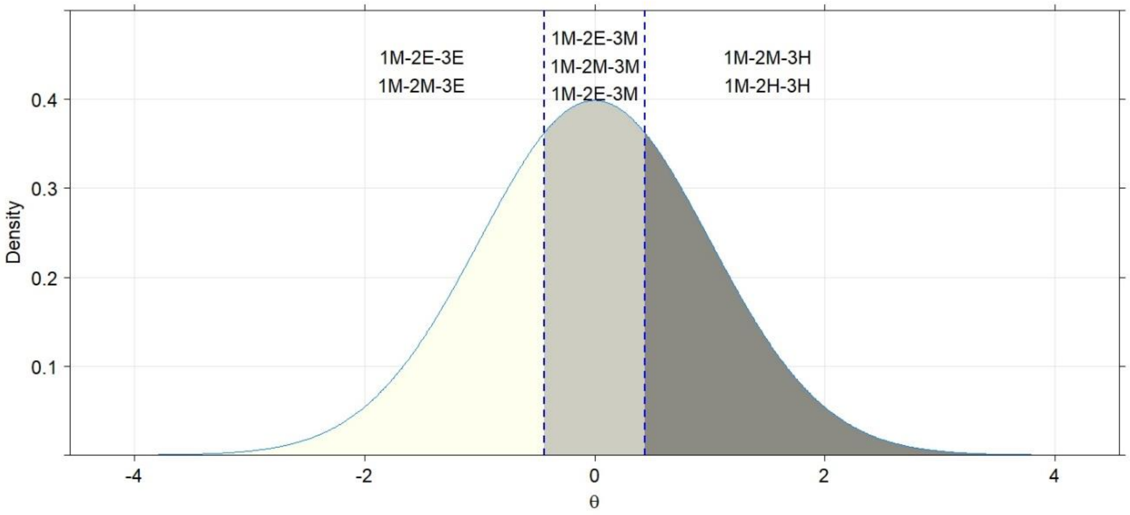

**Figure 9.** An example of route mapping to targeted subpopulations in the 1-3-3 MST proposed by Luo and Kim (2018)

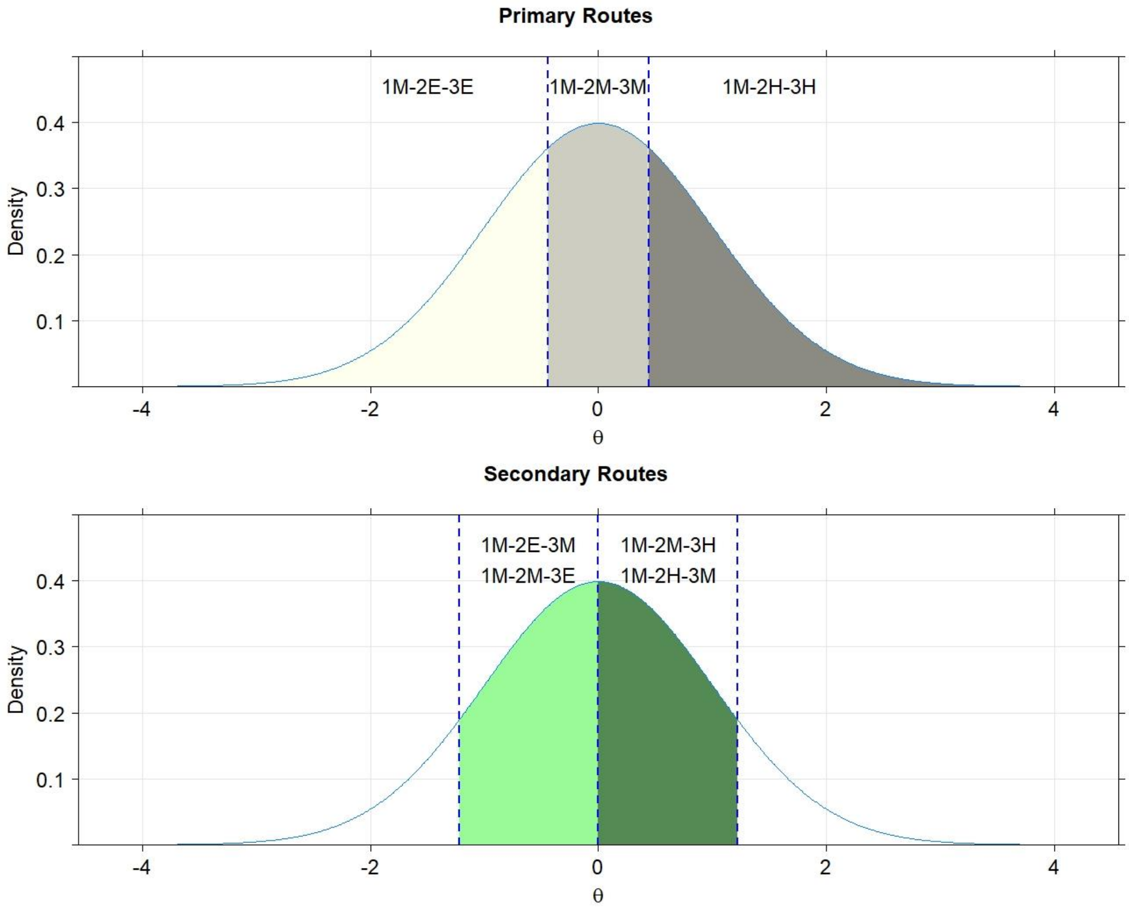

**Figure 10.** An example of modified route mapping to targeted subpopulations in the 1-3-3 MST

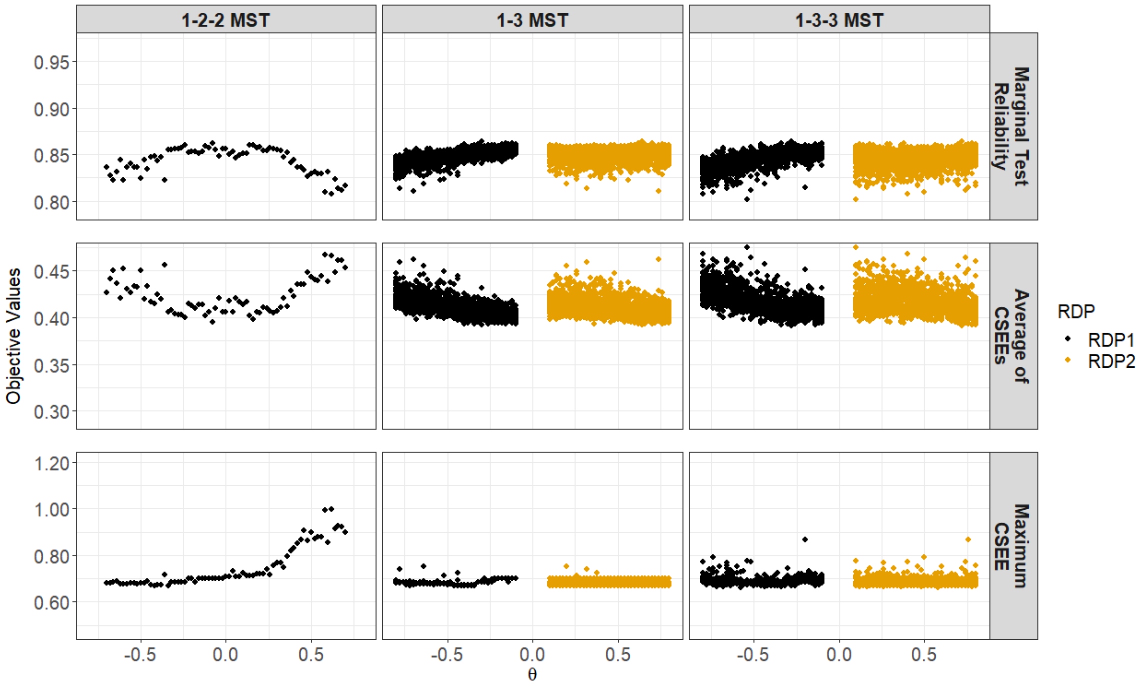

**Figure 11.** Scatter plots between the routing decision points (RDPs) and three objective functions: 32-Item Test under 200-Item Pool

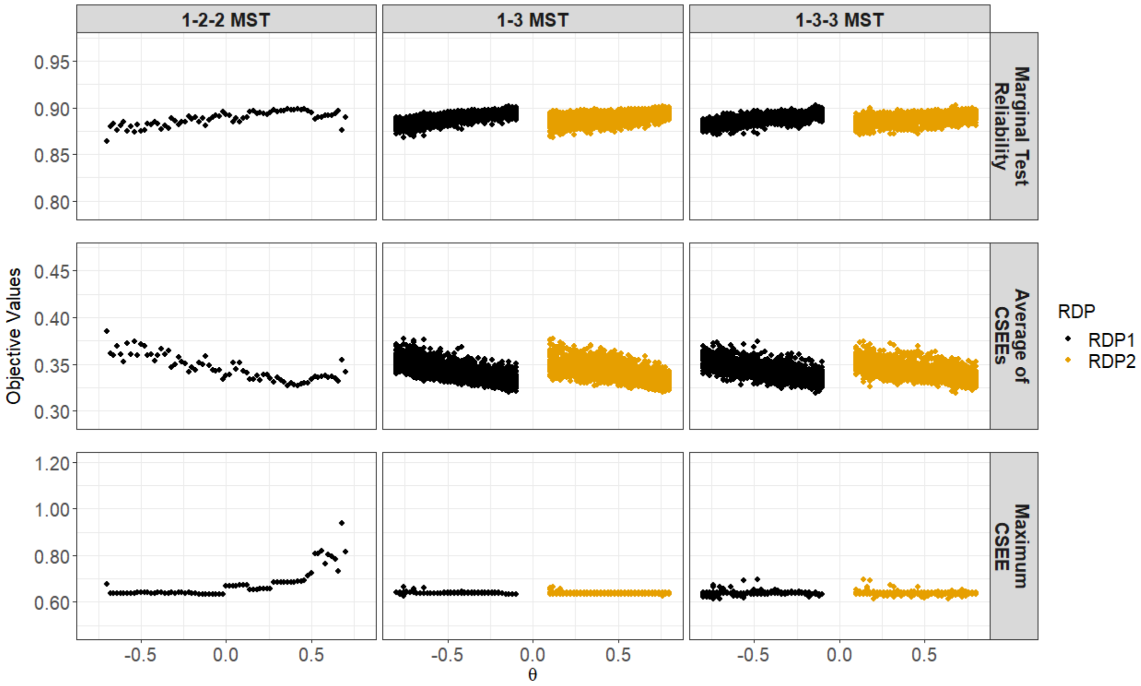

**Figure 12.** Scatter plots between the routing decision points (RDPs) and three objective functions: 60-Item Test under 200-Item Pool

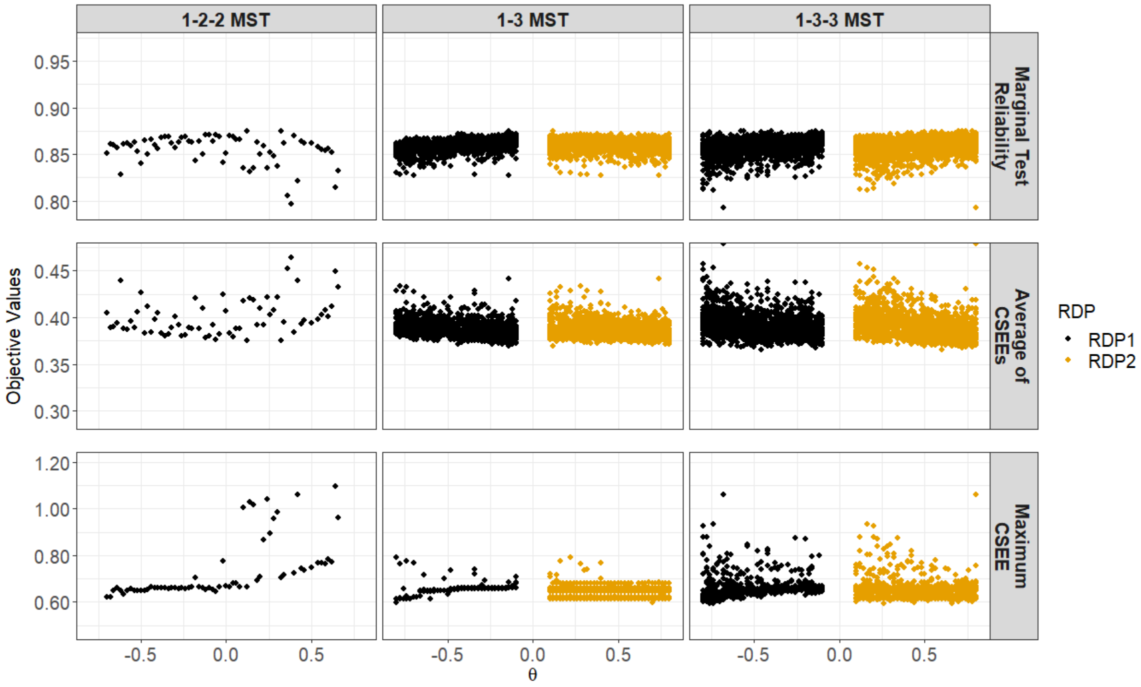

**Figure 13.** Scatter plots between the routing decision points (RDPs) and three objective functions: 32-Item Test under 400-Item Pool

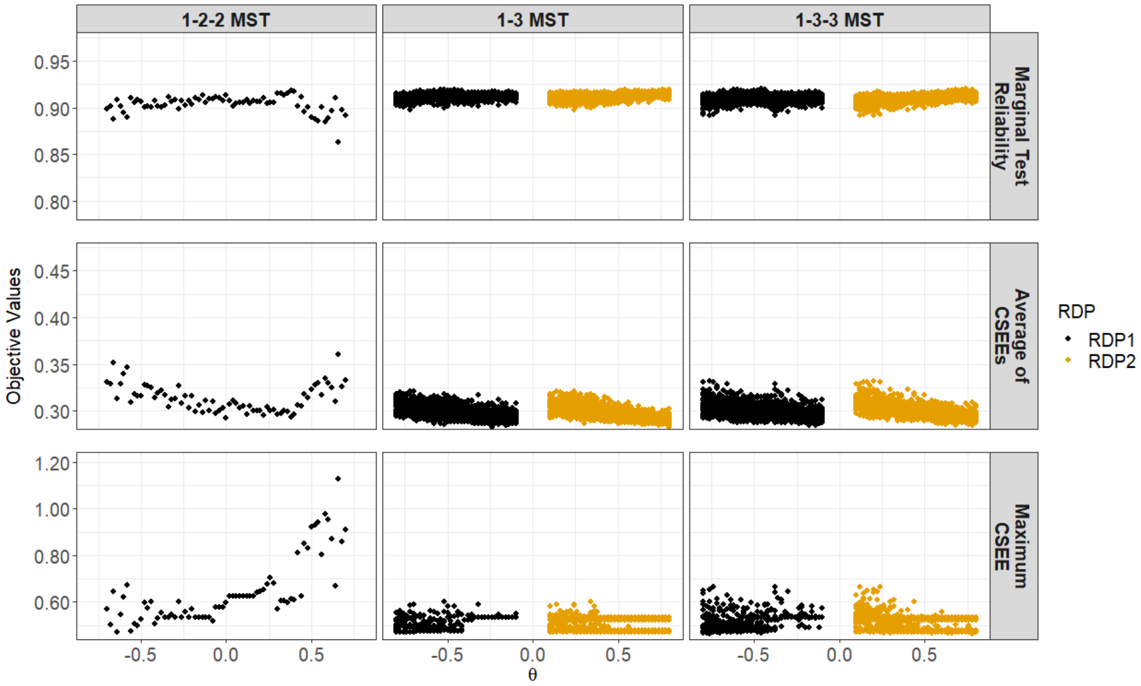

**Figure 14.** Scatter plots between the routing decision points (RDPs) and three objective functions: 60-Item Test under 400-Item Pool

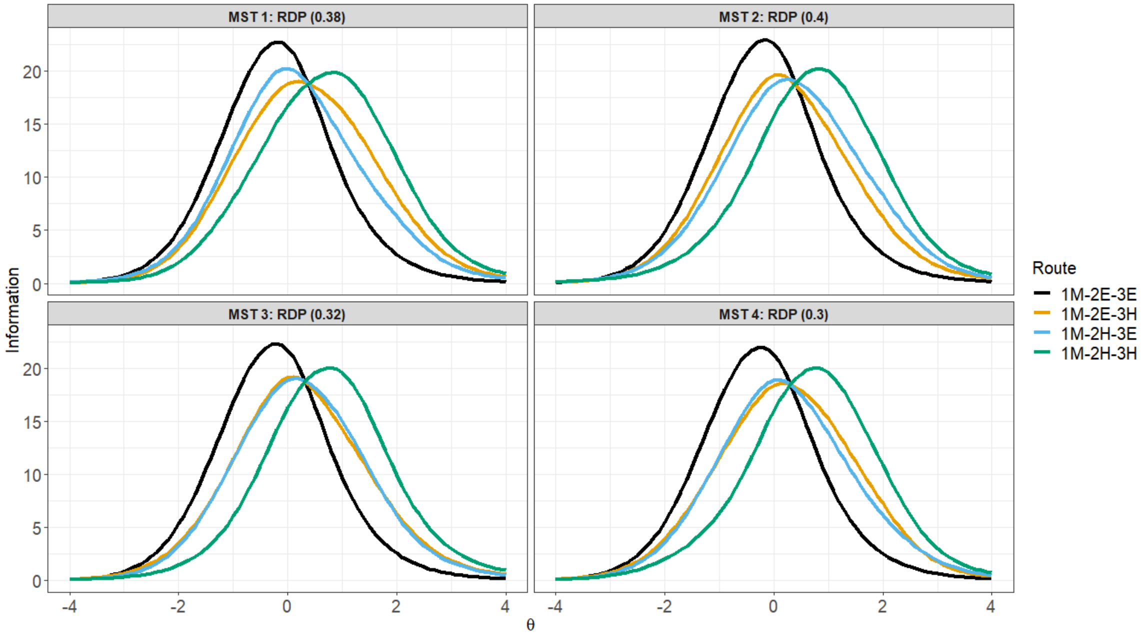

**Figure 15.** Route information functions for the top four MSTs based on marginal test reliability: 1-2-2 MST with 60 Items under 400-Item Pool

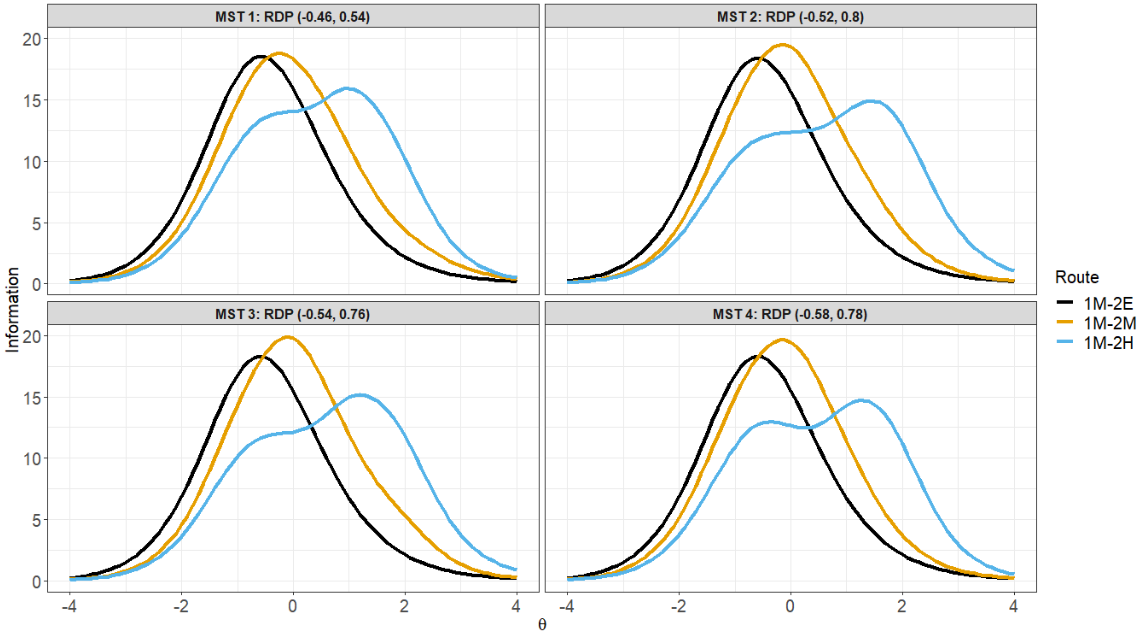

**Figure 16.** Route information functions for the top four MSTs based on marginal test reliability: 1-3 MST with 60 Items under 400-Item Pool

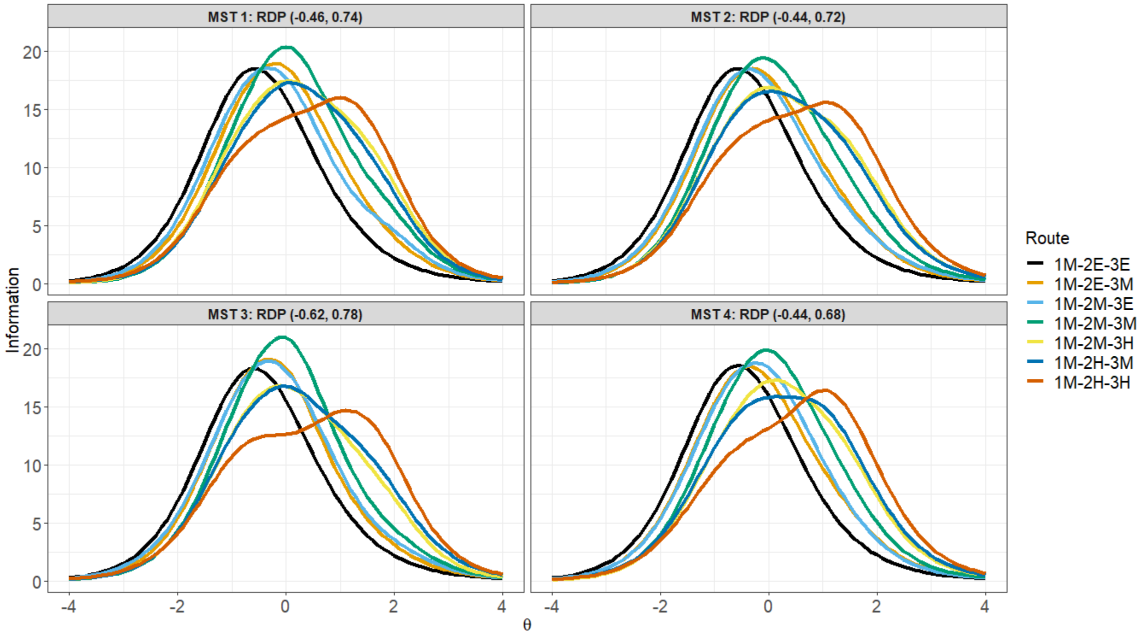

**Figure 17.** Route information functions for the top four MSTs based on marginal test reliability: 1-3-3 MST with 60 Items under 400-Item Pool

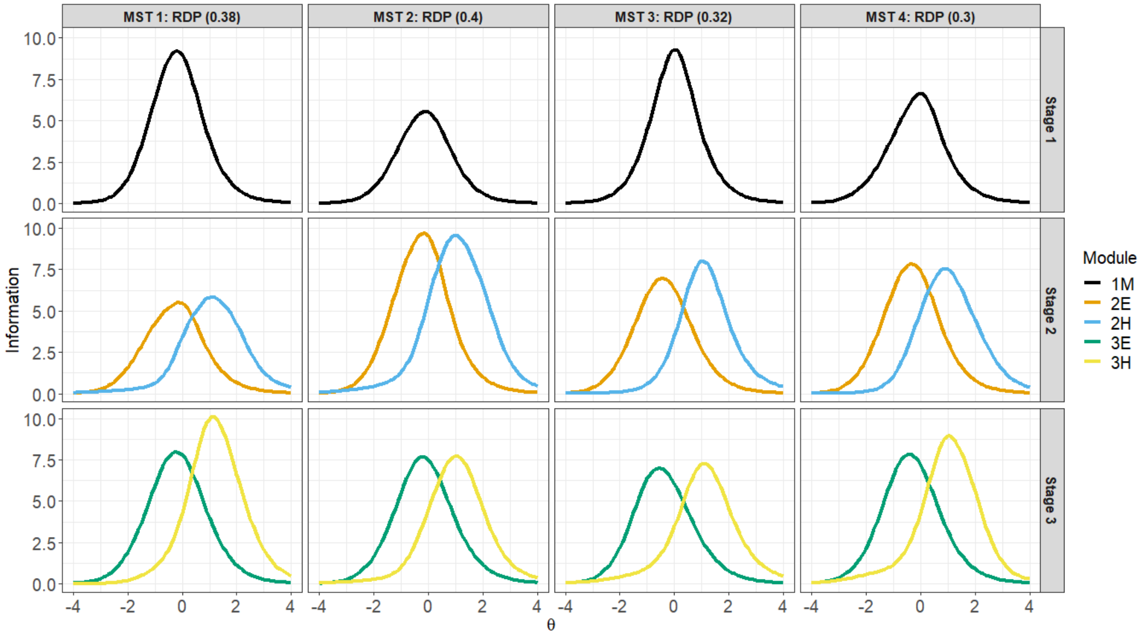

**Figure 18.** Module information functions for the top four MSTs based on marginal test reliability: 1-2-2 MST with 60 Items under 400-Item Pool

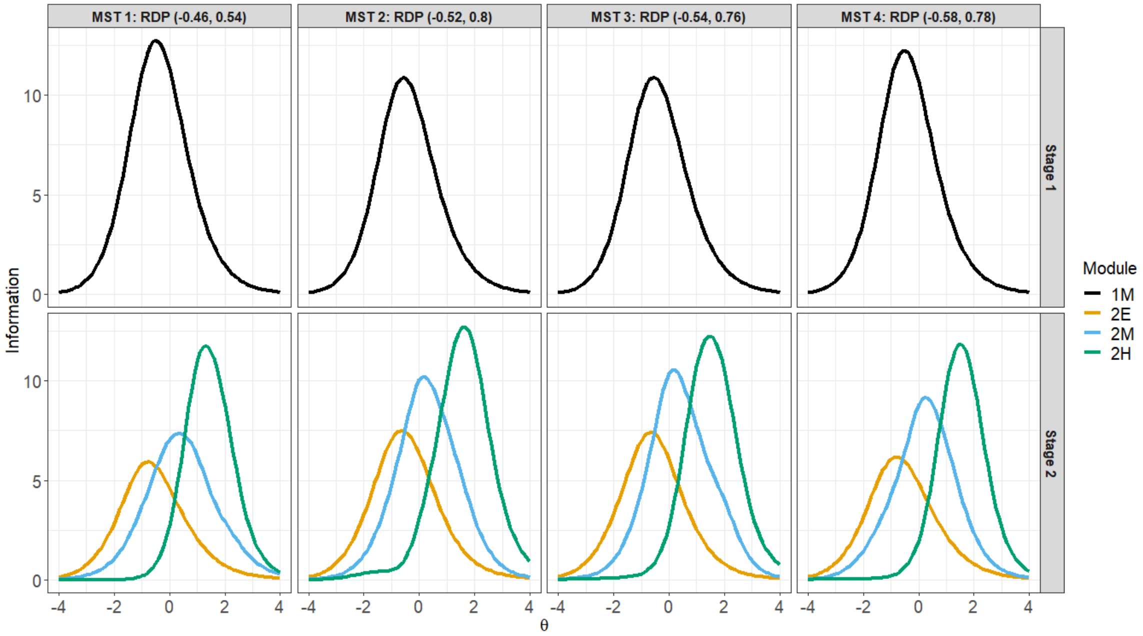

**Figure 19.** Module information functions for the top four MSTs based on marginal test reliability: 1-3 MST with 60 Items under 400-Item Pool

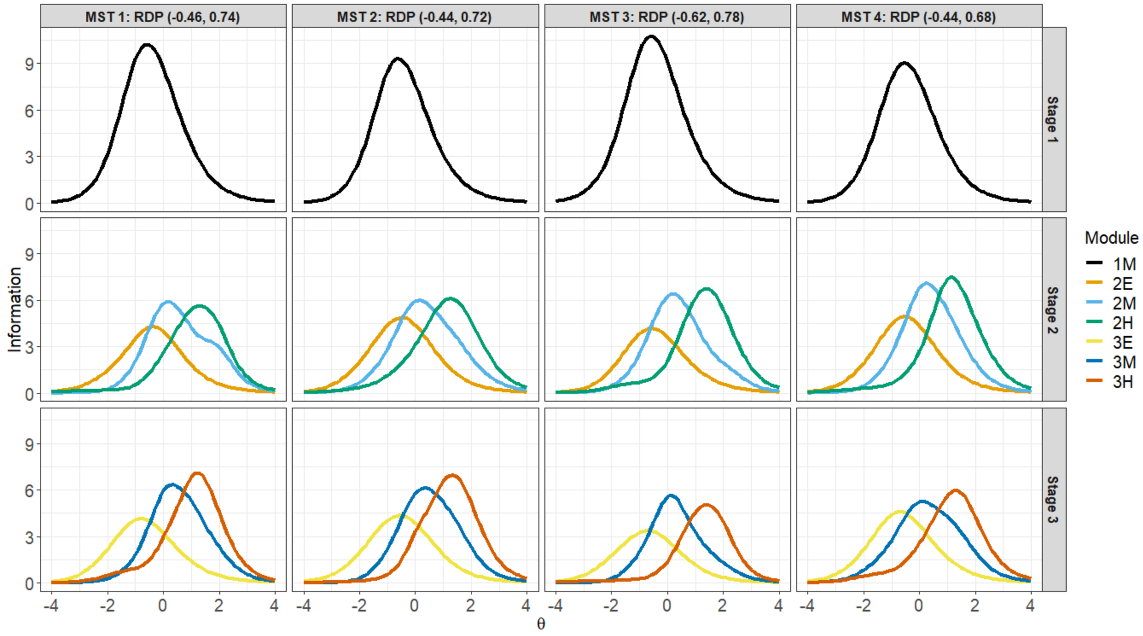

**Figure 20.** Module information functions for the top four MSTs based on marginal test reliability: 1-3-3 MST with 60 Items under 400-Item Pool

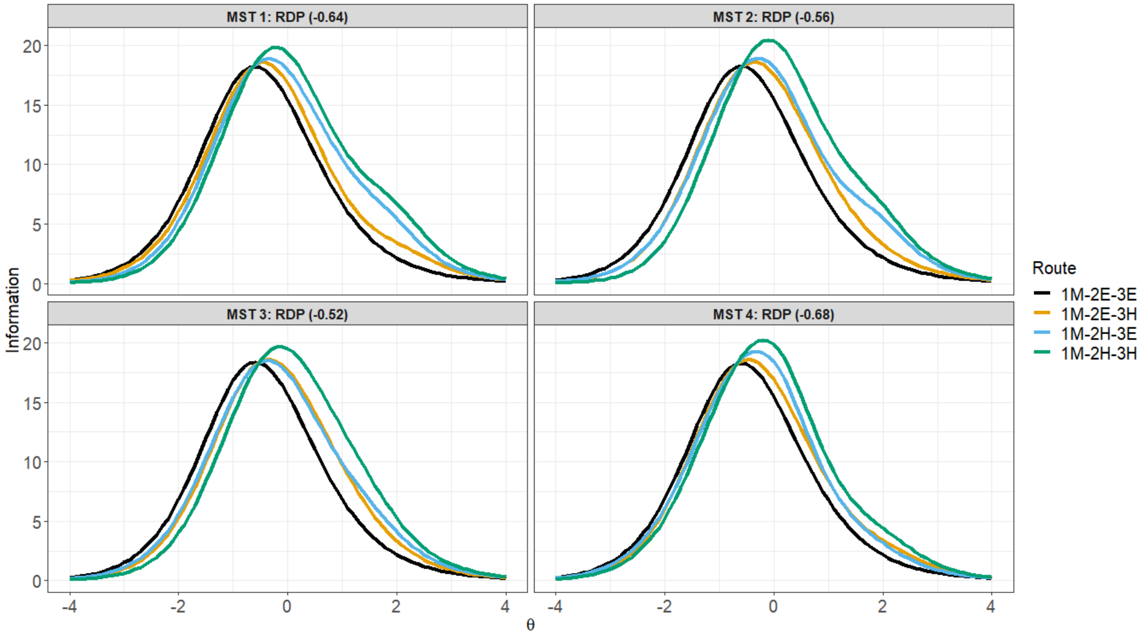

**Figure 21.** Route information functions for the top four MSTs based on maximum CSEE: 1-2-2 MST with 60 Items under 400-Item Pool

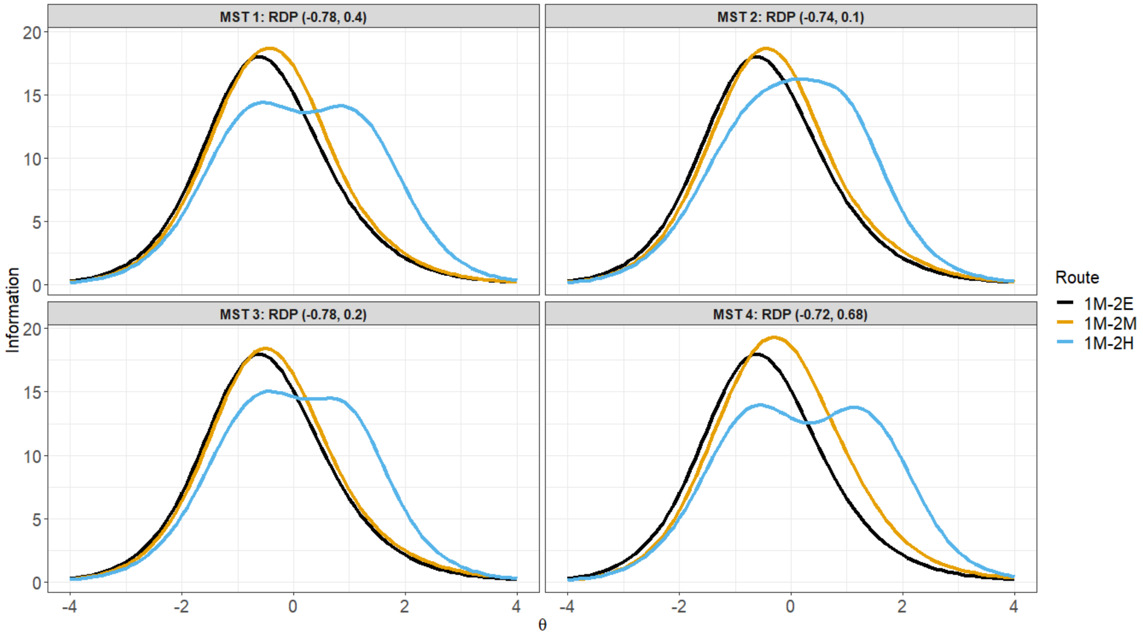

**Figure 22.** Route information functions for the top four MSTs based on maximum CSEE: 1-3 MST with 60 Items under 400-Item Pool

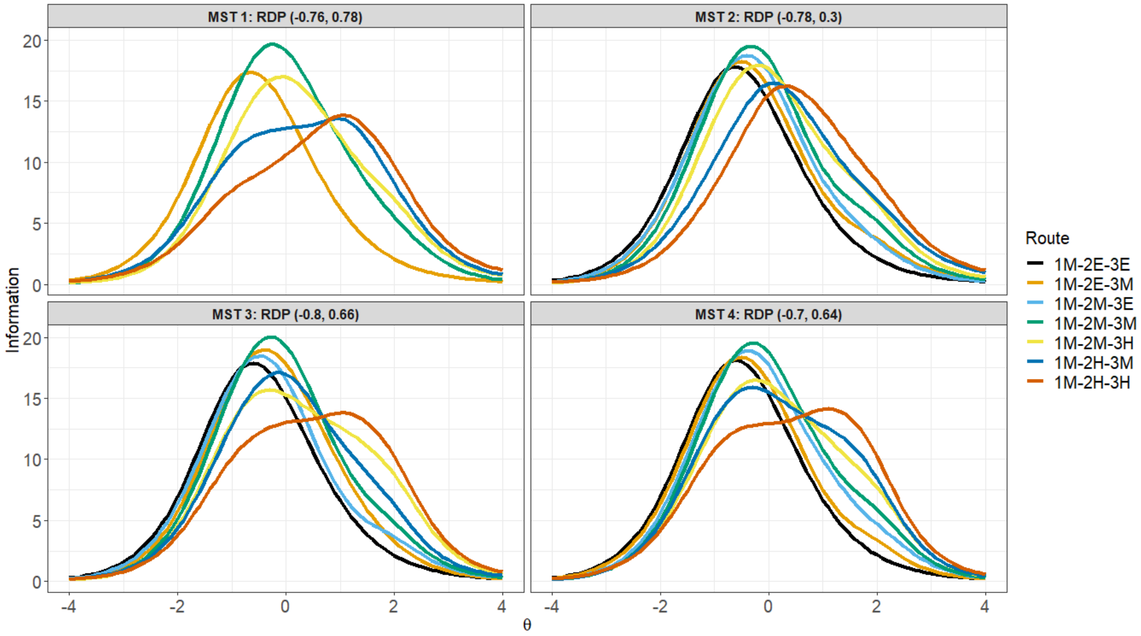

**Figure 23.** Route information functions for the top four MSTs based on maximum CSEE: 1-3-3 MST with 60 Items under 400-Item Pool

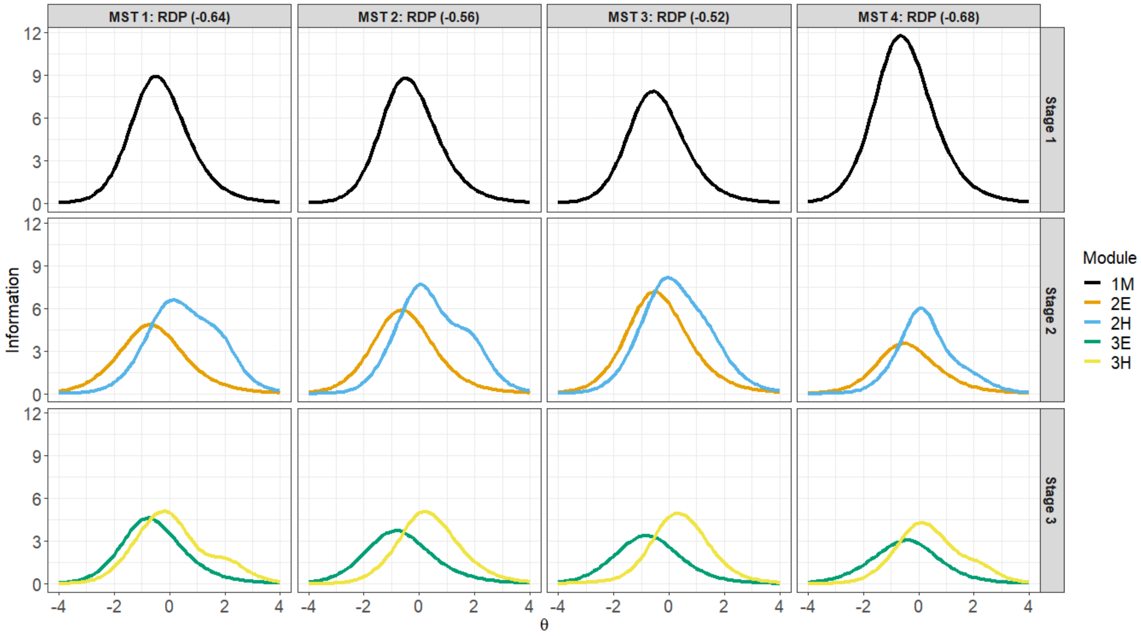

**Figure 24.** Module information functions for the top four MSTs based on maximum CSEE: 1-2-2 MST with 60 Items under 400-Item Pool

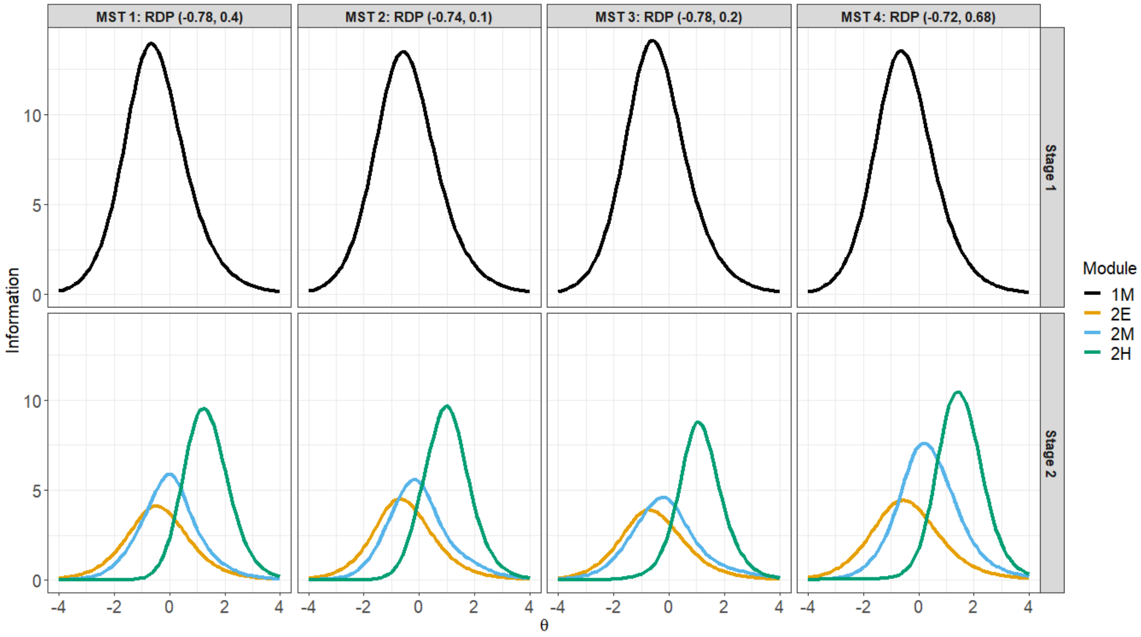

**Figure 25.** Module information functions for the top four MSTs based on maximum CSEE: 1-3 MST with 60 Items under 400-Item Pool

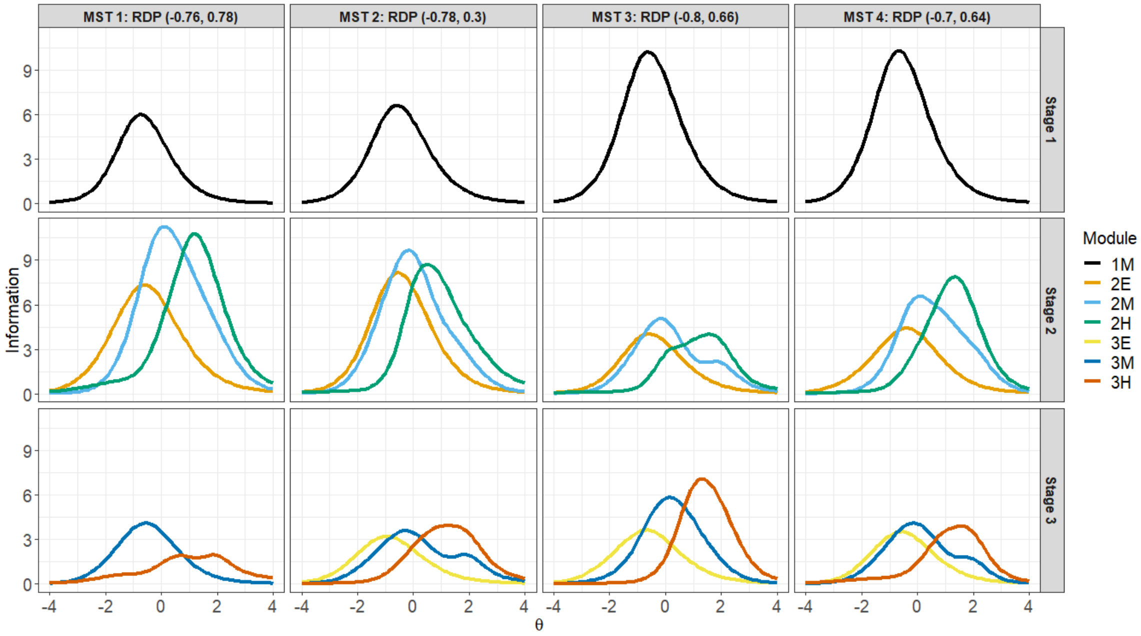

**Figure 26.** Module information functions for the top four MSTs based on maximum CSEE: 1-3-3 MST with 60 Items under 400-Item Pool

# REFERENCES

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

American Educational Research Association, America Psychological Association, & National Council for Measurement in Education. (2014). *Standards for educational and psychological testing.* Washington, DC: Americal Educational Research Association.

Armstrong, R. D., & Roussos, L. (2005). *A method to determine targets for multi-stage adaptive tests. RR-02-07.* Newton, PA: Law School Admission Council.

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.

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

Breithaupt, K., Ariel, A. A., & Hare, D. R. (2010). Assembling an inventory of multistage adaptive testing system. In W. J. van der Linden, & C. A. Glas, *Elements of adaptive testing* (pp. 247-266). New York, NY: Springer.

Breithaupt, K., Ariel, A., & Veldkamp, B. P. (2005). Automated simultaneous assembly for multistage testing. *International Journal of Testing, 5*(3), (pp. 319-330).

Chen, W. H., & Thissen, D. (1999). Estimation of item parameters for the three-parameter logistic model using the marginal likelihood of summed scores. *British Journal of Mathematical and Statistical Psychology, 52*(1), 19-37.

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

Choi, S. W., Grady, M. W., & Dodd, B. G. (2011). A new stopping rule for computerized adaptive testing. *Educational and Psychological Measurement, 71*(1), 37-53.

Dallas, A. (2014). The effects of routing and scoring within a computer adaptive multi-stage framework (unpublished doctoroal dissertation). Greensboro: University of North Carolina.

Diao, Q., & Ren, H. (2018). Constructing Shadow Tests in Variable-Length Adaptive Testing. *Applied psychological measurement, 42*(7), 538-552.

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

Dodd, B. G., Koch, W. R., & De Ayala, R. J. (1993). Computerized adaptive testing using the partial credit model: Effects of item pool characteristics and different stopping rules. *Educational and psychological measurement, 53*(1), 61-77.

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.

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

Han, K. T., & Guo, F. (2014). Multistage testing by shaping modules on the fly. In D. Yan, A. A. von Davier, & C. Lewis, *Computerized Multistage Testing: Theory and applications* (pp. 119-133). Boca Raton, FL: CRC Press.

Harwell , M., Stone, C. A., Hsu, T. C., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement, 20*(2), 101-125.

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

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, H., & Plake, B. S. (1993). Monte Carlo simulation comparison of two-stage testing and computerized adaptive testing. *Annuall meeting of the National Council on Measurement in Education.* Atlanta, GA.

Kim, J., Chung, H., Dodd, B., & 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 Researach Methods, 45*(4), 1087-1089.

Kim, S., & Moses, T. (2014). An investigation of the impact of misrouting under two-stage multistage testing: A simulation study. ETS RR-14-01.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices.* New York, NJ: Springer.

Kolen, M. J., & Tong, Y. (2010). Psychometric properties of IRT proficiency estimates. *Educational Measurement: Issues and Practice, 29*(3), 8-14.

Konis, K. (2009). lpSolveAPI, version 5.5.0.20 [Computer software]. *Retrieved from http://CRAN.R-project.org/package = lpSolveAPI*.

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

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

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true score and equipercentile observed score equatings. *Applied Psychological Measurement, 8*(4), 453-461.

Luecht, R. (2014). Design and implementation of large-scale multistage testing system. In D. Yan, A. A. von Davier, & C. Lewis, *Computerized multistage testing: Theory and applications* (pp. 69-83). Boca Raton, FL: CRC Press.

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

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

Luecht, R. M., & Burgin, W. (2003). Test information targeting strategies for adaptive multistage testing design. *Annual meeting of the National Council on Measuement 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., Brumfield, T., & Breithaupt, K. (2006). A testlet assembly design for adaptive multistage tests. *Applied Measuement in Education, 19*(3), 198-202.

Luo, X., & Kim, D. (2018). A top-down approach to designing the computerized adaptive multistage test. *Journal of Educational Measurement, 55*(2), 243-263.

Meijer, R. R., & Nering, M. L. (1997). Trait level estimation for nunfitting response vectors. *Applied Psychological Measurement, 21*(4), 321-336.

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

Park, R., Kim, J., Chung, H., & Dodd, B. G. (2014). Enhancing pool utilization in constructing the multistage test using mixed-format tests. *Applied Psychological Measurement, 38*(4), 268-280.

Park, R., Kim, J., Chung, H., & Dodd, B. G. (2017). The development of MST test information for the prediction of test performance. *Educational and Psychological Measurement, 77*(4), 570-586.

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

Psychometric Society. (1979). Publication policy regarding Monte Carlo studies. *Psychometrika, 44*, 133-134.

R Core Team. (2018). *R: A language and environment for statistical computing.* Vienna, Austria. URL https://www.R-project.org/.: R Foundation for Statistical Computing. .

Rudner, L. M. (2001). Computing the expected proportion of misclassified examinees. *Practical Assessment, Research & Evaluation, 7*(14), 1-8.

Rudner, L. M. (2005). Expected classification accuracy. *Practical Assessment, Research & Evaluation, 10*(13), 1-4.

Schnipke, D. L., & Reese, L. M. (1997). A comparison of testlet-based designs for computerized adaptive testing. *Annual Meeting of the Americal Educational Research Association.* Chicago, IL.

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

Stocking, M. L. (1996). An alternative method for scoring adaptive tests. *Journal of Educational and Behavioral Statistics, 21*(4), 365-389.

Stocking, M. L., Steffen, M., & Eignor, D. R. (2002). *An exploration of potentially problematic adaptive tests.* ETS RR-02-05.

Swanson, L., & Stocking, M. L. (1993). A model and heuristic for solving very larage item selection problems. *Applied Psychological Measurement, 17*(2), 151-166.

Thissen, D. & Orlando, M. (2001). Item response theory for items scored in two categories. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp.73-140). Mahwah, NJ: Lawrence Erlbaum.

Thissen, D., Pommerich, M., Billeaud, K., & Willams, V. S. (1995). Item response theory for scores on tests including polytomous items with ordered responses. *Applied Psychological Measurement*, *19*(1), 39-49.

van der Linden, W. J. (2000). Optimal assembly of tests with item sets. *Applied Psychological Measurement*, *24*(3), 225-240.

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

Van der Linden, W. J., & Adema, J. J. (1998). Simultaneous assembly of multiple test forms. Journal of *educational measurement, 35*(3), 185-198.

Wang, K. (2017). A fair comparison of the performance of computerized adaptive testing and multistage adaptive testing (unpublished doctoral dissertation). Michigan State University.

Wang, X., Fluegge, L., & Luecht, R. (2012). A large-scale comparative study of the accuracy and efficiency of ca-MST panel design configurations. *Annual meeting of the National Council on Measuement in Education.* Vancouver, BC.

Weissman, A., Belov, D. I., & Armstrong, R. D. (2007). *Information-based versus number-correct routing in multistage classification tests.* Newtown, PA: Law School Admission Council.

Wendler, C., & Bridgeman, B. (2014). *The research foundataion for the GRE revised test: A compendium of studies.* Princeton, NJ: Educational Testing Service.

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

Yan, D., Lewis, C., & von Davier, A. A. (2014). Overview of computerized multistage tests. In D. Yan, A. A. von Davier, & C. Lewis, *Computerized Multistage Testing: Theory and application* (pp. 3-20). Boca Raton, FL: CRC Press.

Yen, W. M. (1984). Obtaining maximum likelihood trait estiamtes from number-correct scores for the three-parameter logistic model. *Journal of Educational Measurement, 21*(2), 93-111.

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). Amherst: Unversity of Massachusetts.

Zenisky, A. L., & Hambleton, R. K. (2014). Multistage test designs: Moving research resutls into practice. In D. Yan, A. A. von Davier, & C. Lewis, *Computerized Multistage Testing: Theory and applications* (pp. 21-37). Boca Raton, FL: CRC Press.

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, *Elements of adaptive testing* (pp. 355-372). New York, NY: Springer.

Zenisky, A. L., Sireci, S. G., Lewis, J., Lim, H., O'Donnell, F., Wells, C. S., Padellaro, F., Jung, H. J.. Pham, D., Hong, S. E., Park, Y., Botha, S., Lee, M., & Garcia, A. (2018). *Massachusetts Adult Proficiency Tests for College and Career Readiness: Technical manual.* Amherst, MA: Center for Educational Assessment: Center for Educational Assessment research report No. 974.

Zheng, Y., & Chang, H. H. (2015). On-the-fly assembled multistage adaptive testing. *Applied Psychological Measurement, 39*(2), 104-118.

Zheng, Y., Nozawa, Y., Gao, X., & Chang, H. H. (2012). Multistage adaptive testing for a large-scale classification test: The designs, automated heuristic assembly, and comparison with other testing modes. ACT RR-2012-6.

Zheng, Y., Wang, C., Culbertson, M. J., & Chang, H. H. (2014). Overview of test assembly methods in multistage testing. In D. Yan, A. A. von Davier, & C. Lewis, *Computerized Multistage Testing: Theory and applictions* (pp. 87-99). Boca Raton, FL: CRC Press.