THE EFFECTS OF TEST SPEEDEDNESS CONTROL WITHIN A COMPUTERIZED ADAPTIVE MULTI-STAGE FRAMEWORK

A DISSERTATION
SUMMITTED TO THE FACULTY OF THE
UNIVERSITY OF MINNESOTA

BY

Qinjun Wang

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

ADVISOR
Ernest C. Davenport, Jr.

March 2019

© 2019
Qinjun Wang

ALL RIGHTS RESERVED

Acknowledgements

This work would not have been possible without the consistent support and guidance of my advisor, Dr. Ernest C. Davenport, who has always found the time to discuss questions and ideas. Over the past five years, Dr. Davenport provided a highly flexible research environment for me to search for the research topic that truly interests me. He not only supported my dissertation writing throughout the entire process, but also taught me to think critically in conducting research in education measurement.

I am grateful to all of my committee members who were more than generous with their expertise and time. Dr. Michael C. Rodriguez has been a great mentor over the past five years. It was his influence from the introductory measurement course that made me take the first step into the world of educational measurement. Working with him in the IGDI projects was one of my most rewarding experiences in graduate school. Dr. Mark L. Davison brought me to the world of Item Response Theory and provided me with the first opportunity to conduct psychometric simulation. It is hard to fully express my appreciation for the support and guidance that Dr. Davison has given me in my early academic career. Dr. Chun Wang has been an amazing role model as a researcher who generously shared her time and expertise to help me to improve my dissertation study. I genuinely appreciate her persistent encouragement and support.

Besides my committee members, I would like to thank my supervisors in IGDI projects, Dr. Alisha Wackerle-Hollman and Dr. Scott R. McConnell, without whom this thesis would not have been possible. They provided endless support for my research studies over the past five years and I would not have learned about adaptive testing and

response time if I had not joined the IGDI Lab. It has been a wonderful experience to work with them.

Nobody has been more important to me in the pursuit of the degree than the members of my family. I would like to thank my supportive parents, whose love and guidance are with me throughout the past five years. Most importantly, I wish to thank my loving wife, Sisi Chen, who gives unending love and inspiration throughout my graduate career.

Abstract

This dissertation examined the overall effects of test speededness control within a computer adaptive multistage (MST) framework. Although the multistage testing framework has become increasingly popular in the testing industry due to its administration efficiency as compared to traditional paper-pencil tests and its quality-control features over item level computerized adaptive testing (CAT), the paucity of research on the effect of the MST routing adaptation on examinees' performance has hampered attempts to fully fulfill the potential of the MST framework. Considering the real time influence of the routing adaptation on examinees who receive negative feedback on their responses, these low ability examinees are more likely to experience a speeded test administration compared to their high ability counterparts. This study proposed a panel configuration design that controls the test speededness level by adding parallel modules in the test via the test assembly phase. By varying both item pool characteristics and MST configuration characteristics, the results from this study showed that the proposed speededness control panel configuration produces significantly better ability estimation compared to traditional MST designs without considering test speededness.

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

1.1 An Overview of Relevant Research in Educational Testing

Powered by the instant computation capability of modern computers, computer-based testing (CBT) is becoming widely accepted in military selection and placement, psychological assessment, and educational assessment. Although there are many types of CBT delivery models, most can be differentiated by three test administration characteristics: (1) test assembly process, that is, whether the tests are assembled in real time or tests are pre-assembled; (2) whether the test length is fixed across test takers or variable; and (3) whether the test is adaptive to test takers' responses. Computerized adaptive testing (CAT), which possesses the testing characteristics of real-time assembly, variable test length, and the item level adaptation, is typically designed as a test delivery mode to make a test more efficient. By targeting the difficulty of each administered item to the provisional ability estimation and updating the score estimate after each item in real time, the CAT can improve the stability of score estimation over a non-adaptive test of the same length or maintain a prescribed score estimation precision with a shorter test.

In spite of the elevated estimating precision on person ability and more flexible test scheduling and assembly, CAT still can be improved in operational settings considering some of its shortcomings, such as violation of local independence assumption, lack of control over item ordering and potential context effects, lack of opportunity to review items, and overwhelming data management and computation demands (Hambleton, Swaminathan, & Rogers, 1991; Vispoel, 1998, and Yen, 1993).

Preassembled linear fixed tests, on the other hand, are non-adaptive and fixed in length where every examinee sees a common collection of items once they are assigned to a test form. Compared with CAT, linear fixed tests may lose in efficiency by being non-adaptive, but they outperform the CAT in providing several crucial quality control and quality assurance advantages, such as the capability to review and change the item set for the test form.

Multistage Testing (MST), a test delivery mode that combines the strengths of both linear form test and CAT, receives steady support in operational testing practices (Luecht & Nungester, 1998). In MST, test adaptation takes place between item sets or testlets based on the cumulative performance on previous item sets, rather than between each item. As a consequence, MST features a distinct streamlined assembly process from CAT. Selected items are grouped into a *module* (testlets) by their psychometric characteristics, and modules are grouped into a *panel*, which will be assigned randomly to a test taker. Within each panel, the MST program configures the selected modules in the predetermined multistage structure where the positioning of each module in the structure relies on its overall difficulty. Thus, MST blends the quality control / quality assurance capabilities of linear form tests and the efficiency of CAT.

The genesis of the modern computer is not only a catalyzer for the transformation of educational testing from linear to adaptive, but also the facilitator of response time analysis in educational test settings. Response times have been considered as a meaningful dependent variable in cognitive psychology since the mid 1950s (Luce, 1986). It has been widely accepted that the time it takes someone to process a task

reflects how the person processed the task. However, despite the history of response time research in experimental cognitive psychology, the practice of response time research in educational testing is restricted due to the difficulties in recording response times in real world test settings. Early research methods used in collecting response times, such as having test takers record their start and stop times for each item, are not only intrusive, but compromises the standard test administration protocol (Rindler, 1979). With the advent of computer-based testing (CBT), response times can be collected effortlessly and unobtrusively in standard operational test settings. Researchers in educational measurement are in a much better state to start making use of response times to address various time-relevant issues in educational tests.

Tremendous strides have been made in response time studies in educational tests. Different aspects of response time conceptualization, such as differentiation between response time and speed (e.g., van der Linden, 2009), speed and accuracy relationship (e.g., Scrams & Schnipke, 1997), and time limit effect (e.g., Hopkins, 1998) were investigated. Furthermore, various quantitative methods were developed to model response time data (e.g., Scheiblechner, 1979; Tatsuoka, 1980; Thissen, 1983; Roskam 1997) and to conduct post hoc response time analyses. Particularly, the post hoc response time analyses include aberrant behavior detection (e.g., Marianti, Fox, Avetisyan, Veldkamp, & Tijmstra, 2014), subgroup differences in testing time conception (Jenkins & Holmes, 1999), test taker motivation assessment (Kong & Wise, 2005), test strategy implementation (Schnipke & Scrams, 1997). With the collaboration between psychometricians and cognitive psychologists, the current spectrum of response time

studies has extended to complex assessment, diagnostic feedback, and construct validity (Scrams & Schnipke, 2002).

As a natural progression of advancing educational testing with additional administration information, the use of response times was considered in computerized testing while it posed both challenges and potentials in CAT development. Individualized item sets are tailored for test takers in different ability groups while items with different time intensities can be included in a testlet and administered to test takers. Therefore, it is reasonable to assume that differential speededness can occur in a CAT administration. The test might become speeded for some test takers while other test takers might experience a CAT that is more similar to a power test (e.g., van der Linden, 2006). Consequently, additional constraints are warranted to prevent this phenomenon during CAT assembly. Despite the challenges in test construction, employing response time enables the researchers to tune the CAT to satisfy both psychometric and security purposes (Veldkamp, 2016). Item level response time variation, for instance, can be utilized when the construct of interest is the ability to solve the task within a reasonable amount of time. Similar to the post hoc analyses in fixed format tests where aberrant behavior detection can be conducted with the item level and test level response time data, response time pattern can also be examined after the CAT administration for potential cheating behavior (van der Linden, 2008).

Akin to the implementation of response time in CAT, availability of response time information on computers also sparked explorations of the potentials of response time in elevating the MST program performance. Van der Linden, Breithaupt, Chuah, and

Zhang (2007) presented a differential speededness examination method in the framework of MST. To detect systematic differences in time intensity between items in challenging testlets and easier testlets, the paper suggested plotting the distributions of the estimated time-intensity parameters for items in the testlets as a function of the difficulty level of the testlets. In addition, in order to investigate test taker behavior during MST, Lee and Jia (2014) proposed a five-step procedure to identify rapid-guessing behavior which comprises defining response time threshold, classifying behaviors, and three validity checks.

1.2 Purpose and Rationale

While the existing abundant literature on response time implementation in CAT indicates a smooth process of migrating response time analyses in MST since both test modes share the adaptation characteristics, it seems current literatures have not investigated some unique yet unneglectable structural properties in MST that could have posed both challenges and opportunities to the development of response time implementation in MST. Despite the psychometric similarity in test delivery, the administrative structure of MST can be differentiated from its counterpart in CAT via two aspects: 1) tests are partitioned into several stages while each between-stage routing may be accompanied by a perceptible test difficulty change while the item level difficulty adaptations are subtle in CAT. 2) an additional source of test speededness could be introduced to a certain group during the MST administration due to real time performance feedback that examinees receive after each routing procedure. Taking these two MST-specific attributes into account when implementing response time

consideration in MST design warrants both specification and enhancement of existing testing panel configuration in MST context.

Nothing carries more weight in MST construction than producing equated test administration for all test takers before the onset of the test. To secure test consistency and fairness across all test takers while taking advantage of the ease in response time collection, the present study introduces a new perspective on MST panel configuration by using a response time related variable, test speededness, to advance and elevate the MST model performance. In particular, a new test panel configuration is proposed to control the varied real time psychological influences of MST routing adaptations on examinees. By tuning the item selection for specific modules using item level time characteristics in certain routing pathways, the proposed MST design is able to control unbalanced test speededness due to the routing adaptation. Using an R-programmed simulation, the estimating precision of the proposed test panel configuration is evaluated by comparing it to a traditional MST design without test speededness consideration. In addition, the study also shows the adverse impact of uncontrolled test speededness on ability estimation precision by varying the levels of in-test speeded response patterns.

Furthermore, this study has a practical focus, ideally providing testing agencies with empirical evidence, or at least suggesting psychometrically sound research methods, to examine their own MST administrations with an emphasis on the responses of examinees who are being routed to relatively easier modules and the consistency of their updated ability estimations throughout the administration. In addition, the study explores the minimal item pool requirement based on the proposed test design and assembly

demands. Based on the results, testing agencies could invest in the necessary item writing activities to improve the item supply over time. Although this study uses simulated data, the basic distributions of item parameters are simulated based on a large-scale, operational item pool. Lastly, this study also includes manipulations of the item pool characteristics and the in-test speeded response patterns that are not commonly employed in operational practices, but the goal of these examinations is to investigate the potential impact of rather extreme item pool conditions and uncommon response behaviors on overall MST model performance. In that respect, we not only evaluate the technical viability of the proposed panel configuration design, but also examine practical conditions that could be problematic in operational uses.

CHAPTER II

REVIEW OF THE LITERATURE

Computer-based testing (CBT) fuels the exploration of the response time studies in educational testing, which benefits researchers' understanding of examinees, as well as the item (Marianti, Fox, Avetisyan, Veldkamp, & Tijmstra, 2014). While the importance and practicality of response time in CBT is well-represented in the current literatures, uncharted territory emerged from the combination of response time and Multistage Testing. This chapter summarizes the relevant topics necessary to understand the design and outcome of the present study. The first section introduces three main concepts related to the response time research in educational testing contexts and the concept of test speededness. The second section delineates the response time implementations in Computerized Adaptive Testing. The third describes the major components in Multistage Testing development with an emphasis on the automated test assembly algorithm. The final section reviews the current studies on test anxiety in adaptive tests, including the Multistage Testing mode, and discusses the key aspects of test anxiety in MST administration that are being omitted as well as the possibility of advancing MST model performance.

2.1 Response Times in Educational Testing

The term response time (RT) in educational testing refers to the time a test taker spends on an item in a test. Schnipke and Scrams (2002) provided a comprehensive review of the RT studies that were conducted before 2000. The majority of the RT models included in the review were crafted for simple and automatic cognitive tasks, or

in a more generic term, speed tests (Gulliksen, 1950). Gulliksen dichotomized examinations as power tests and speed tests. Specifically, a power test is a test with unlimited time but a fixed number of items representing a range of difficulties and all examinees are expected to finish the test while a speed test administers a much larger number of items within a limited time and completing the entire test is not required. These varied test focuses between two test types result in two unique measuring goals. The purpose of the power test is to measure how accurately examinees respond to items while the speed test results describe examinee's cognitive processing speed through 1) examining the total time used to complete a fixed number of items or 2) examining the total count of completed items within a fixed time interval. However, major barriers arise when administering both tests since the test designs rarely align with the administrative purposes of educational assessments.

Therefore, the ideas of power test and speed test appear to be unrealistic in educational assessments (van der Linden & Hambleton, 1997) since the majority of the power test-oriented educational tests impose time limits for practical reasons. As noted by van Breukelen (2005), the RT models for current time-limit tests are addressing both power and speed aspects of the tests simultaneously. Consequently, both responses and RTs have to be taken into account in order to understand the response process in a time-limited test.

2.1.1 Response Time Conceptualization

Admittedly, one can opt for a simpler perspective and use the collected item responses to judge the test takers' abilities or the quality of items. However, consider the following example of two students, Alex and Bill, who both took a two-item statistics test

on linear regression as an in-class quiz. The test results show that Alex answered both items correctly, but Bill only answered the first item correctly. When solely considering their responses, we might conclude that Alex understands the tested regression knowledge better than Bill. But after examining the collected response time information, we noticed that Bill took 55 seconds to answer both items while Alex only used 10 seconds, which is unreasonably fast since reading the whole item stem should take a longer time. After being questioned, Alex admitted he guessed both items.

This straightforward yet realistic illustration shows that a valuable source of information would have been missed if we ignore the response time information. In the following three sections, we review three major facets of the relation between response times and test responses to facilitate the understanding of response time in educational tests.

Test Taker's Ability and Speed

Test takers' knowledge about a specific domain is not observable, test stakeholders gain information through the manifestations of test takers' knowledge on an administered test or questionnaire. Item Response Theory (IRT) has been developed to fulfill the need of making inferences about unobserved abilities of test takers using their observed responses. IRT models define the probability of a correct item level response from a test taker as a function of test taker's ability and item level characteristics. A unidimensional 2-parameter logistic (2PL) model, for instance, has an ability parameter for all test takers and two parameters, representing difficulty and discrimination, for the

items (Lord & Novick, 1968; Embretson & Reise, 2000). The normal-ogive formulation of the 2PL IRT model is expressed as

$$E(\mathrm{Y}) = \Phi(\alpha\theta - \beta), \tag{2.1}$$

Where $E(\mathrm{Y})$ denotes the probability of providing a correct answer (Y = 1) to the item given the test taker's ability, θ. $\Phi(\bullet)$ represents the normal cumulative distribution function. β and α denote the difficulty and discrimination of the item, respectively. As indicated by the formula, ability level is positively related to probability of providing a correct answer.

Given the test uses collected responses to measure test taker's ability as the underlying construct, it could also be assumed to measure test taker's speed as the unobservable construct through the collected response times. Therefore, the notion of speed in a response time model can be described as follows: a test taker with higher processing speed can perform in a lower expected response time. Response speed, a test taker specific parameter similar to person ability, can be integrated into a model for response times that accounts for person level differences.

Efforts in applying response speed in computerized educational measurement date back to the 1970s, Scheiblechner (1979) used separate person speed and item speed parameters to capture the distribution of response time in an exponential function. Tatsuoka (1980) conceptualized the response speed parameter as the expected response time for the given test taker over an infinite set of items with similar characteristics. Thissen (1983) and Verhelst (1997) considered the response speed and response accuracy simultaneously in power tests and speed tests, respectively. More recently, van der

Linden (2006) proposed the basic assumption that the person operates at constant ability and speed during the administration. Test takers do not have the freedom to choose their abilities as well as the speed levels in a test while the ability and speed were structured as a speed-accuracy tradeoff relation within individual. The response times thereby become conditionally independent given speed once the test taker decided to apply either a speed strategy or accuracy strategy.

The notion of a speed-accuracy tradeoff is well represented in psychological reaction-time research. Particularly, it represents a positive relation between the proportion of correct responses and the average time on the items (Luce, 1986). Considering the time spent on an item is determined by the test taker's speed and the person specific ability parameter dictates the response correctness, the counterpart of speed-accuracy tradeoff in an educational test context is therefore a speed-ability tradeoff (van der Linden, 2009). Van der Linden also noted that the tradeoff is entirely a within-person phenomenon and a given test taker has control of speed during a test but has to take the consequence of fluctuated response correctness due to the speed change. For two test takers with differing abilities, it is intuitive that the more abled test taker can achieve a higher rate of correctness than the other if both test takers process the item at the same speed.

Despite the speed-ability tradeoff, the "effective ability parameter" is another vital component that is closely related to the relationship between speed and ability. Roskam (1997) defined the effective ability parameter as the product of mental speed and processing time. In Roskam's model, the traditional ability parameter of the 1PL model

was replaced by "effective ability". Roskam (1997) assumed that on the person level, the rate of correctness depends on effective ability, which increases as more time is spent on an item. The rate of this increase is characterized by the person parameter, mental speed. With an exponential scale, the effective ability becomes the sum of mental speed and response time. This results in the model as follows:

$$p_i(\theta_j) = \{1 + \exp[-(\theta_j + \ln t_{ij} - b_i)]\}^{-1}. \quad (2.2)$$

The product of mental speed and time is represented by the $\theta_j + \ln t_{ij}$, while the model also follows the speed-accuracy tradeoff that the discrepancy between ability and item difficulty can be compensated by increased response time usage. Van der Linden (2009) provided an updated explanation of the effective ability parameter by describing the term as the certain response speed level chosen by an individual based on his/her perception of the test time limit and judgement of the difficulty of current and unanswered items. He further noted that test takers' decisions on setting the effective ability parameter may vary across items. Therefore, the estimated test scores will not reflect their true abilities since the test takers actively alter the response speed during the test. A test fairness issue would emerge when different test forms are involved, and some test takers will be forced to elevate their response speed to finish the test.

Item Difficulty and Time Intensity

Time differences not only exists between test takers' response behaviors, but between items too. Items in a test usually vary in their difficulties and person ability can be obtained by examining responses to items with various difficulties. Extending the same rationale to the response time model, items should also vary in the amount of time needed to solve them and it is referred as the item level time intensity (van der Linden,

2009). Therefore, the ability and difficulty relationship in item response models can be paralleled by the speed and time intensity relationship in response time models.

To illustrate the conceptual difference between item difficulty and item time intensity. Consider two arithmetic items (van der Linden, 2009), the first one requires the sum calculation of 315 + 918 and the second requires the calculation of 315 + 918 + 42 + 74. The difficulties of the two items should not depart substantially from each other since being able to perform three-digit number addition indicates adeptness in two-digit number addition. Nevertheless, it is apparent that the second item requires more time as it has a longer series of addition. Thus, the item difficulty refers to the level of cognitive challenge posed by an item, whereas the item time intensity describes the total amount of processing effort demanded by an item.

Research effort in characterizing the item level processing time requirement began with modeling the mixture of item difficulty and time intensity. Thissen (1983) introduced an item effect on the RT parameter along with a person "slowness parameter" in his regression model for response time. Wang & Hanson's model (2005) combines RTs and response speed by extending a traditional 3PL model. Particularly, they introduced a new component, "$-\rho_j d_i / t_{ij}$", to adjust the person ability θ, where the new component produced a collective modification on the ability estimate by taking account of person speed, ρ_j, item intensity, d_i, and response time, t_{ij}. The development of a RT model that specifically describes the time intensity characteristics was not addressed until the advent of the lognormal RT model (van der Linden, 2006) where the difference in time intensity between items are modeled by an item parameter λ. It functions

analogously as the item difficulty parameter in an IRT model. Akin to the difference between difficulty and ability in the IRT model that infers the probability of response correctness, the expected log-RT in the lognormal RT model is indicated by the difference between the personal speed, ζ, and the item intensity, λ.

Local Independence of Response Times

The local independence assumption on responses between different items has a history in IRT (e.g., Lord, 1980; Holland & Rosenbaum, 1986). Typically, IRT modeling treats a test taker as a random draw from a population (Hollan, 1990). With a specified normal population distribution of test takers' ability parameter, the ability parameter is perceived as a random person effect that captures the between-test takers differences. Thus, the ability parameter is supposed to account for all associations between responses on different items.

Extending the local independence concept to RT models, the person speed parameter can also be considered as a random effect across test takers and these random effects model the heterogeneity in the collected person level response times. Consequently, no covariation between RTs across items should remain unexplained after conditioning on the random speed parameter. In both IRT and RT models, the random effects (ability and speed) are assumed to be constant over items given the imposed local independence assumption. In other words, test takers are assumed to perform at a consistent level of ability and speed during a test.

The concept of conditional (local) independence in ability and speed is also the key in explaining the contradicting conclusions on the dependencies between RTs and

responses that can be found from empirical study results. Earlier studies that investigated the relationship between RTs and responses presented substantial correlation between the two outcomes. Bergstrom, Gershon, and Lunz (1994), for instance, analyzed the response data from a computerized adaptive certification exam via a hierarchical linear model and found that examinees spent more time on items they answered incorrectly than on items they answered correctly. The same conclusion was drawn by Hornke (2000), who found a negative relationship between RTs and response outcomes on a prescreening military test. These findings contradict the speed-accuracy tradeoff in reaction-time psychology, where longer time usage invariantly results in higher correctness, as well as the earlier models with the assumption of positive association between RTs and response.

Van der Linden (2009) noted that the inconsistencies on RTs and responses dependency may be the consequence of data aggregation across test takers. Within each item-RT combination, only one RT and one response can be obtained for a given test taker. In other words, no correlation can be calculated. The aforementioned studies that correlated these two outcomes across test takers may result in spurious correlations with hidden covariates unexplained. If the more abled test takers, for instance, work faster, a positive relationship will be observed. If the low performing group work faster, a negative relationship is expected. With local independence of ability and speed in mind, it is assumed that within test takers, the response time and response outcome are conditionally independent on every item. Therefore, it is important to recognize the assumption of local independence when modeling the relationship between RTs and responses. Specifically, they are independent from each other on the same item.

2.1.2 Test Speededness

By tracking test taker's response times to individual items, test stakeholders gain insights of the unobservable test takers' attributes as well as the new characteristics of items that were previously unclear. Furthermore, applying the collected response times to various test-relevant contexts allows test stakeholders and test developers to monitor and tune the test to achieve better estimation precision as well as a more equitable interpretation of the test outcome. As most standardized tests are administered with a specific time limit, test speededness draws increasing attention in the development of the tests that employ varied form, because the uncontrolled test speededness between test forms may lead to a fair proportion of test takers not having sufficient time to complete the test, which would undermine the test's validity (Lu & Sireci, 2007). Historically, speededness was defined via two perspectives, score-oriented and fairness-oriented. Specifically, the score-oriented perspective treats a test as speeded when the final score is determined by the number of items attempted and the response accuracy (Bejar, 1985) while the fairness-oriented perspective defines test speededness as the extent to which some test takers are disadvantaged by the test limit relative to others (Schnipke, 1995). Recently, the definition of speededness was further expanded by van der Linden (2011) as a mixture of three factors: the speed of the test taker, the amount of labor required by an item, and the test time limit.

With the prevalence of computerized adaptive tests in high-stakes tests, appropriate control of speededness became even more challenging due to the fact that items with various time related characteristics (e.g., time intensity) can be selected for different test takers which can easily lead to differential speededness across forms.

Besides the test fairness concerns, speededness can also bring issues in item pretesting and calibration studies (van der Linden, 2011). When the time availability between pretesting condition and operational condition are different, the item estimation will be biased, particularly for items located at the end of a speeded form. Moreover, careful examination of the relationship between test speededness and time limit is necessary when considering time accommodation (Stretch & Osborne, 2005) because improper time extension would introduce an additional source of inequity.

In the context of item response theory, the probability of a correct response is solely driven by the test taker's proficiency and the characteristics of the item (Hambleton, Swaminathan, & Rogers, 1991). If substantial speededness exists in a time limit test while the speed factor is not properly modeled, the fundamental IRT assumption is violated as the incorrect responses are attributable to either limited proficiency or low response speed. Therefore, unaddressed speededness obscures the interpretation of the ability estimate as well as the item level statistics (Oshima, 1994). Oshima (1994) found that for items located at the end of the test, the difficulty and discrimination parameters can be overestimated and guessing parameter can be underestimated if the test speededness factor was not captured in a speeded form. In the investigation of the dependency of item position and item statistics, Wise (1986) found significant relationship between item characteristics and item position for both speeded and nonspeeded forms in the Medical College Admission Test (MCAT), which implies the violation of local independence. Zeniksy, Hambleton, and Sireci (2002) further examined the verbal section of the MCAT and concluded that rapid guessing and nonresponse was more likely at the end of the test which inflates the item difficulty estimate.

IRT based methods for assessing speededness date back to the 1980s. Through the speededness investigation of the Test of English as a Foreign Language (TOEFL), Bejar (1985) proposed two indices, the item-level index and the examinee-level index, to identify the model data misfit due to the presence of speededness. Yamamoto (1995) introduced the extended HYBRID model that takes account of both ability-based response and guessing-based response to assess the degree of test speededness as well as the speededness influence on ability and item parameter estimation. Using the estimated proportion of test takers who switch to guessing at each item, the model was capable of gathering more accurate description of speededness by conditioning the probabilities on test taker's ability and strategy use. Cohen, Wollack, Bolt and Morch (2002) developed a mixture Rasch model where they allowed for item difficulty parameters to vary across groups of test takers who produced speeded response and their peers who didn't experience test speededness. In its application, the model found the difficulty of items shifted higher at the end of the test for the speeded group, since the test takers were believed to perform less well compared with their nonspeeded counterparts. The study also found that the test takers who experienced test speededness appeared to perform better on the test compared with their peers in speededness control condition, but only on items located at the beginning and the middle of the test.

While the speededness of a test is manifested by test takers not finishing all items within the time limit, item-level response times naturally became an indicator for speededness since responding to an item with considerably less time than what is required by a thorough consideration of an item signals the phenomenon that the test

taker answered the item in haste, hence the existence of speededness. Using response times along with response accuracy, researchers are in a better position to detect test speededness. Schnipke (1995) utilized item level response times on an analytical section from the Graduate Record Examinations (GRE) to investigate the test taker's speed change. She found that the typical speeded test taker usually started the test at a slower speed and abruptly switched to rapid-guessing strategy at a certain point of the test. On the item level, previous findings of items appearing later in the test are more likely to be time intensive were confirmed in her study. Particularly, those items involved significantly more short response times and incorrect responses.

Furthermore, Schnipke and Scrams (1997) used response times in a two-state mixture model with an assumption of unidirectional switch from solution behavior to rapid-guessing behavior, that is, they considered test takers would only switch from the accuracy-oriented strategy to the speed-oriented strategy and would not focus on response accuracy after employing speed strategy. The model was further extended to a more generic form (Wise & DeMars, 2006) by removing the restriction on unidirectional strategy-taking route during the test. Given the estimated item time characteristics and person speed parameter from the aforementioned lognormal model, van der Linden (2011) quantified test speededness as the risk, π, of running out of time which is a function of the test taker's speed, item time intensity and time discrimination, and the time limit. The risk, π, can fill various roles in understanding test speededness either before or after the test. For instance, it can be applied to the empirical check on a subjective claim that test takers running out of time before answering all items is due to the speeded test, it also serves the pretest speeded risk evaluation before the test

administration. With a simple transformation, the time limit requirement for a given test can be calculated with a predetermined level of risk, π. A complete introduction to the methods for risk calculation and test limit design is explained in the next chapter.

2.2 Implementing Response Time in Computerized Adaptive Tests

The benefits of computer-based test delivery are twofold. One advantage is the feasibility of adapting the test difficulty to the ability level of test takers during the test administration. Another advantage is the capability of recording not only the response outcome on each item, but the amount of time the test taker spent in considering and answering each item. In the past few decades, various methods of selecting appropriate items for individual test takers were developed by leveraging the combined advantages from computer-based tests.

2.2.1 Item Selection Algorithms Using Response Time

To date, the majority of response time applications in CAT cluster around two areas. One application constrains the item selection with respect to response times to control for differential speededness. Another cluster of algorithms strives for minimizing the test time in conjunction with maximizing ability estimation accuracy.

As mentioned in section 2.1.2, ignoring the time intensity characteristics of items can be problematic in CAT as it permits differential speededness. This phenomenon is indicative of unfairness between test takers since one finished the test under constant high time pressure whereas his or her peers might answer all items with ample time. Van der Linden, Scrams, and Schnipke (1999) introduced a response time-based constraint in the

item selection procedure that drives the test under a consistent test speededness blueprint. The concept takes the mathematical form of

$$\sum_{i \in R_{g-1}} t_{ij} x_i + \sum_{i \in I/R_{g-1}} E\left[t_{ij}\right] x_i \leq t_{tot}, \tag{2.3}$$

where the sum of time usage, t_{ij}, on administered items, $i \in R_{g-1}$, and the expected time usage on the remaining items given person *j* is restricted to be lower than the total test time. In addition, the set I/R_{g-1} represents the set of items remaining after g – 1items have been administered.

In practice, the most informative item often requires a large amount of time to answer, it is considered a more efficient strategy to use multiple less time-consuming items with lower information within the same time slot. Similarly, constructing a CAT administration that requires the minimized test time while selecting relatively informative items is also a common administrative goal. The maximum information per time unit (MICT) method was proposed to achieve this goal (Fan, Wang, Chang, & Douglas, 2012). The criterion function for selecting the *g*th item can be formulated as

$$\max_{i \in I/R_{g-1}} \left\{ \frac{\mathrm{I}_i\left(\hat{\theta}_j\right)}{E\left[t_{ij} \middle| \hat{\tau}_j\right]} \right\}, \tag{2.4}$$

where the $\hat{\theta}_j$ and $\hat{\tau}_j$ denote the current ability and speededness estimates. The time required to answer the item *i* can be calculated via the log normal model based on the item time intensity and the estimated test taker speed. This method outperformed other item selection algorithm candidates in minimizing the total test time (Veldkamp, 2016).

Choe & Kern (2014) broaden the MICT method by proposing a generalized time-weighted maximum information criterion (GMICT). It is expressed as follows:

$$\mathrm{IT}_i^G = \frac{I_i(\hat{\theta}_j)}{\left|\mathrm{E}(t_{ij}|\hat{\tau}_j) - v\right|^w}. \tag{2.5}$$

Two new parameters, a centering value (v) and an exponent weight (w), were added in the original MICT model. The centering value v facilitates the speededness control since it boosts the selection probability for items whose expected response time is close to the center of time intensities across all items. The exponent weighting parameter handles how the departure of expected time from center is adjusted. It is also obvious that the GMICT reduced to the MICT model when $v = 0$ and $w = 1$, and it can be further simplified to maximum information criterion when $w = 0$. As w increases, GMICT places more weight on absolute deviation of expected time from v than the maximum information criterion does, that is, items are chosen with increasing emphasis on minimizing response time and it is also accompanied by a worsened estimation accuracy.

2.2.2 Other Response Time Applications in CAT

Although the response times are primarily used for constructing accurate person ability estimation and maintaining invariant test speededness across test takers, they contain valuable information for various aspects in the CAT context, such as post hoc test analyses and item pool development. Two studies are reviewed to indicate the versatility of response time in CAT administration.

Aberrant Behavior Detection

Van der Linden (2003) introduced a residual analysis-based method for examining the aberrant behavior of examinees who obtain preknowledge of the CAT item

pool. Using the item level response time data, it is argued that knowing part of the item pool would result in unexpected correct response and sizeable decrease in item level response time. In general, applying response times in aberrant testing behaviors produces two technical advantages in statistical analyses over using item responses exclusively: 1) the continuous nature of response time data resolves the barrier of converting dichotomous item response residuals to an asymptotically normally distributed variable; 2) Since the item selection algorithms in CAT aim to minimize the difference between the examinee's ability and the selected item's difficulty, the expected probability of correct response on a given item would center around 0.5 throughout the CAT administration. Therefore, the response residuals, which is the difference between the response outcomes (0s and 1s) and the expected correct response probability, would be equivalently located on values of -0.5 and 0.5. However, the continuous nature of response times is not constrained by this dichotomous residual distribution pattern.

Both classical (large-sample approximations) and Bayesian (posterior prediction) methods were employed in the simulation. The performance evaluation of both methods is comprised of a detection rate (proportion of simulees that were correctly flagged as aberrant with respect to the total number of simulees with aberrant behavior) and a false-alarm rate (proportion of simulees that were erroneously flagged). The classical check method showed a false-alarm rate of 0.05 and a detection rate of 0.3 for examinees who showed high aberrance level in the simulated test-taking behaviors. It was also found that the test length was not a significant factor on the results. The Bayesian check method, on the other hand, doubled the detection rate in the classical method with a substantial increase in false-alarms as well. Despite the investigated item pool preknowledge

behavior, other types of test taking aberrances in CAT remain to be investigated such as warming-up effect, instruction misunderstanding, and aggregated fatigue toward the end of the test, while it is also suggested to consider these checks on response times as supportive evidence for aberrance supposition instead of decisive ones.

CAT Item Pool Rapid-Guessing Threshold Setting

Wise and Ma (2012) presented a Normative Threshold (NT) method for identifying non-effortful responding behavior to replace the common three-second threshold, which is considered too conservative for items that contains substantial text reading and items that are mentally taxing. Acknowledging the fact that a rapid guessing response may vary based on the different time demands across items, the NT method defined an item level time threshold through a certain percentage of the item specific average response time with a maximum 10 seconds threshold. If an item requires on average 40 seconds to solve, for instance, a 10 percent threshold (NT10) would be 4 seconds and a 20 percent threshold (NT20) would be 8 second. Generally, higher percentage yields larger proportion of rapid guessing behavior classification than would the common three-second method.

In terms of the accuracy rates for each threshold method, the rapid guesses for the five-option math test should exhibit an accuracy rate of 0.2. Both NT10 and three second threshold produced similar results. As the NT percentage increases to 15 and 20, the observed accuracy rate was substantially elevated above 0.2, which is indicative of effortful responding behavior. Splitting the newly included responses as the transition from NT10 to NT20 into solution responses and guess responses. It was found that 31% of the rapid guesses were correct after the threshold moved from NT10 to NT15 and 47%

of the rapid guesses were correct once the threshold was positioned at NT20. This trend was consistent for both verbal and math tests in the study and the obtained statistics suggested the NT10 maintained finer precision for distinguishing rapid guessing behavior from solution behavior than its other two NT siblings.

2.3 Multistage Testing (MST) and Test Assembly Procedures

2.3.1 MST Introduction

Building upon the item level test adaptation in CAT, multistage testing (MST) is an analogous test delivery mode where the test administration adapts between item sets instead of individual items. In MST terminology, these item sets are often referred to as modules (Luecht & Nungester, 1998) or testlets (Wainer & Kiely, 1987) and can be described as short versions of linear test forms where some pre-specified number of items are administered and analyzed as a unit to meet particular test specifications.

The different approaches in the assignment and selection of modules in creating the entire test administration reflects an important distinction among diverse MST designs and administration strategies. That is, modules can be adaptively selected and administered in real time based on the test taker's response or grouped within the predetermined test administration block called a "panel". By combining multiple particular modules, a panel can meet the explicit statistical targets, as well as specified content and other qualitative test features. Luecht & Burgin (2003) showed that the standardized, consistent panel configuration approach results in several operational advantages, including scoring and data management ease, tighter quality control over test assembly, item selection, and module exposure. In this study, the pre-assembled panels that consist of a group of eligible modules is used.

In a constructed MST panel, test adaptation is realized by routing examinees between testing stages, an administrative division that separates the test into different sections. The number of stages to be implemented is determined by the prespecified test design considertions such as the content coverage level and the measurement precision. Within the stage, at least two modules that differ from each other on the average difficulty are required. In assigning modules within each stage, as with CAT, a test taker receives a module that is targeted in difficulty to the ability estimated from performance on the previous stage to achieve an optimal level of measurement information. As a result, high-performing test takers receive modules of higher average difficulty while relatively easier modules are presented to the less able group. Operationally, MST design is very flexible and the final exact MST program structure and its operational delivery mechanism is a product of a series of considerations that affect the test efficiency. In essence, these considerations can be decomposed to five components: item pool, panel construction, module scoring, routing strategy, and automated test assembly.

2.3.2 MST Item Pool

As a crucial component in the MST measurement framework, an item pool needs to be properly developed to incorporate the particular test content as well as the specified statistical and nonstatistical constraints. A sufficiently large size, item quality, and MST driven design are all decisive aspects that affect the successfulness of the test assembly process. Previous research suggested that adding new items is not adequate to uphold standards of MST item pool quality over time (Ariel, van der Linden, & Veldkamp, 2006).

Breithaupt (2010) noted a natural tension between creating large item pools with low item level exposure and cost reduction, he also noted that this conflict is directly linked with item retirement. Breithaupt argued that the lifespan for case study items should be much shorter because of the fast paced field evolution, whereas items in algebra assessment may be retired at a far less rapid pace. Veldkamp and van der Linden (2010) sought to maintain the item pool with a shadow test approach, which uses traditional item pool blueprints and content blueprints to help avoid content/statistical concerns within a given item pool. Constraints are placed on each module to minimize item cost and test level constraints are implemented to control for test information, number of stimuli, and content coverage etc.

In addition to pool maintenance, another consideration that needs to be addressed is the item pool characteristics that closely relate to the precision of measurement outcome. Jodoin (2006) investigated the degree to which the item discrimination parameter distribution and the item pool coverage of test content affect the measurement precision and classification accuracy. The examined relationship was found to be significant. Wang et al. (2012) found that the quality of the item pool affects the operational efficiency of the panel design. They also indicated that the item pool construction should adapt to panel configuration.

2.3.3 MST Design

An often-cited advantage of test adaptation is the testing time reduction that resulted from the customizable approach in test implementation. While some of the practical test design issues are shared between CAT and MST, there are, however, facets

of MST development that are distinct from CAT to warrant a review of design variables present in MST.

Panel Configuration

In MST, each panel is constructed of multiple modules to be capable of self-adapting an examination form to each test taker, similar to CAT, using performance over a series of two or more testing stages. Panel design configuration in MST, however, can vary in the following ways: 1) the number of stages; 2) the length of each module; 3) the number of modules per stage; and 4) the distribution of item difficulty among modules (Luecht & Burgin, 2003). Broadly speaking, these factors are driven by the purpose of the test, available items, and test specifications.

While the majority of the MST literature revolve around two- and three-stage tests in which everyone is assigned the same number of stages, one of the exceptions, computerized mastery testing (CMT), involves a variable-length mastery test where the number of modules delivered varies on the person level (Luecht, Nungester, & Hadadi, 1996). Choosing the number of stages is usually a policy decision. For instance, stake holders of high-stakes tests may not be inclined to use a two-stage test due to an impression of test takers being unable to recover if their true abilities were not accurately reflected in the first stage and they are fixed at a lower level module. In standardized stage setting, researches showed that three or four stages generally provide increased measurement precision (e.g., Patsula, 1999; Jodoin, 2006, & Hendrickson, 2007).

Considering module length, studies present a wide range of choices from 20 items (e.g., Jodoin, Zenisky, & Hambleton, 2006) to 60 items (Luecht & Nungester, 1998).

Instead of constant module length configuration, researchers have explored various length module arrangements, such as longer first-stage MST (e.g., Xing & Hambleton, 2004) or tests with more items in later stages (e.g., Reese, Schnipke & Luebke, 1999). Patsula (1999) defined two strategies as the routing test strategy and the higher stage strategy where the rationale for the first strategy is improving measurement accuracy in the first stage before routing while administering more items in the later stage to capitalize on the information collected after routing has been conducted. Nevertheless, he showed that varying number of items within each module did not affect the ability estimation precision.

Furthermore, in terms of the number of modules in each stage, its design is related to several aspects, including level of specified routing precision, the depth and width of the item pool, and the degree of overlap between modules. Most MST programs use one module at stage one. Lord (1971) and Kim & Plake (1993) found the measurement accuracy is affected by stage two module count. Armstrong et al. (2004) indicated that three modules are sufficiently capable to produce desired ability estimates.

Once the first stage has been delivered, another issue is the relative difficulty of modules in each of the subsequent stages. The difficulty level of the module is targeted at a specific range on the IRT ability scale. For instance, constructing two levels within a stage may involve targeting the two modules at 0.0 and 0.5 (e.g., Breithaupt & Hare, 2007). Targeting module difficulty generally means locating the average *b*-parameters of items in the given module and such averages can be obtained by either peaked or

nonpeaked distribution (Lord, 1980). Specifically, peaked modules contain items of similar difficulty while its nonpeaked counterpart is made of items with varied difficulty.

Routing and Scoring Strategies

In MST, routing and scoring is operated individually by each specific route. A high ability test taker in a 1-3 panel configuration design, for instance, might take the first (medium difficulty) module in stage 1 and then be adaptively routed to the difficult module on the second stage. Test takers are routed through the stages to individual modules targeted in average difficulty to their estimated proficiency. Thus, routing in MST is the synonym of adapting in CAT. Among all the possible routes in a MST panel design, the primary routes are the ones representing the expected routes taken by most test takers who perform with constant proficiency level while the auxiliary routes capture the recovery routes for test takers close to the routing cut point. The possible routing pathway considerations are usually a policy issue as the number of possible pathways for routing is flexible and can be controlled by the testing program (Luecht & Nungester, 1998). In most of the MST programs, if two adjacent stages have at least three stages each, routing pathways between the easiest module in one stage to the most difficult module in the other stage are usually prohibited since such a drastic change in ability update is largely due to measurement error.

Two routing methods have been widely used for assigning test takers to the next well-matched module. The first method is number correct (NC) score based (Luecht, Brumfield & Breithaupt, 2006) that uses IRT estimates to develop NC cut scores for the panel. Essentially, test developers obtain the NC scores for the pre-specified cuts and use

those NC decisions to route examinees through the panel. The NC decision point scores can be calculated by summing the probability of a correct response across items using any IRT model given the item parameters. As a variation of NC score routing, Jodoin (2006) introduced the defined population interval strategy (DPI), a routing method that forces a predetermined proportion of test takers to take particular routes. This is accomplished by finding the ability points that match with the specific percentiles of the cumulative distribution. As an example of this, consider a 1-3 panel design, in order to place equal number of test takers at each of three modules in stage 2, the scores of the 33rd and 67th percentiles would be the routing point. These two NC based strategies differ in the operational emphasis of the test program, the first method focuses on efficiency and precision by placing the test taker to the route that is maximally informative while the other one prioritizes the equalization of module exposure.

The second routing method is IRT based and it is designed to match the provisional ability estimates at each of the routing point with module difficulty for maximized efficiency (e.g., Kim & Plake, 1993; Jodoin, 2003 and Hambleton & Xing, 2006). Person scores have to be calculated at the end of each module in order to make routing decisions while the responses from previous administered modules have to be considered. Using the summed Test Information Functions (TIF) of previous modules and each of the alternative modules in the next stage, the routing point can be decided by matching the target TIF with the alternative TIF sums.

Once the routing rule is decided, two types of scoring rules can be set up, the number correct scoring and the IRT-based proficiency estimate. Computational ease is

one advantage of using number correct scoring as the test program no longer has to perform calibrations on the fly to route test takers and only requires IRT at the end for final score estimation. The other scoring method uses IRT-based person score, which is usually computed by maximum likelihood estimation (MLE) or expected a posteriori (EAP) estimation. Compared with NC scoring, test takers would be given more credit for correctly answering items that are either highly discriminating or very difficult.

2.3.4 Automated Test Assembly

Generally, the test assembly in MST incorporates statistical and non-statistical test specifications simultaneously with mathematical algorithms to select items and construct multiple modules and panels (Zenisky & Hambleton, 2014). The complexity of the assembly task and the volume of work load in most large-scale testing contexts require computerized and psychometrically sound automated test assembly (ATA) algorithms. While the ATA methods for MST were built on the extensive psychometric research of CAT item selection and test assembly approaches, particular challenging issues for ATA implementation in MST are many. As described by Luecht & Nungester (1998), Luecht, Brumfield & Breithaupt (2006) and van der Linden (2005), the challenges include item bank maintenance, the requirement of having the algorithm meet an object function (e.g., Test Information Function), the varied specification for different modules, and the module security under multiple replication of panel configuration.

Luecht and Nungester (1998) proposed two heuristic assembly strategies, “top-down” and “bottom-up”, for panel configuration. The top-down strategy utilizes modules that are not completely parallel in panel configuration to achieve test-level specifications and constraints. In other words, modules are not exchangeable between panels since the

test specifications are not equivalently loaded across modules. The bottom-up strategy, on the other hand, assembles parallel modules that met common module level specification such as statistical target and content features. Building multiple equivalent versions of each module allows for the mixed-and-matched procedure to create multiple permutations of the panels.

Under MST, all ATA algorithms serve the same goal, that is, creating panels and modules to consistently match multiple statistical targets (Luecht, 1998). The most widely used target in the MST assembly is the Test Information Function (TIF), which denotes the amount of measurement precision that can be achieved in various regions of the latent proficiency scale. As argued by Luecht & Burgin (2003), TIF was chosen to fulfill three tasks: 1) to guarantee measurement precision; 2) to derive targets for simultaneous module production; and 3) to control conditional exposure of items. In practice, using one or more TIFs specified by test developers, the ATA algorithms aim to choose items from the bank so that the difference between target TIF and empirical TIF approaches zero for individual modules or for combinations of modules.

With respect to ATA method research, one promising approach which has been implemented in operational MST is the Normalized Weighted Absolute Deviations Heuristic (NWADH; e.g., Luecht & Nungester, 1998; Patsula, 1999; Hambleton & Xing, 2006), which uses item-level information functions to manage the utility and availability of items to construct modules under specified constraints. The normalization procedure allows for the simultaneous realization of numerous objective functions. The absolute deviation denotes the absolute difference between the target test information function and

the current function. The weighing procedure assigns weights to items within content areas that require administration, which prioritizes the content constraints over statistical targets. Compared with other existing algorithms, NWADH generates solutions that meets all constraints with rapid convergence while it is not designed to seek the optimized solution. However, considering computational expense and real time application, testing programs usually prefer the NWADH over the other algorithms that seek the best solution.

Another prevalent ATA method in module and panel generation is the Linear Programming method (LP; van der Linden & Adema, 1998) where specified constraints (e.g., content specifications and item exposure rate) can be strictly satisfied using the inequality based mathematical optimization procedure. Van der Linden (2005) proposed a Mixed-Integer Programming (MIP) method, which comprised of both integer and continuous decision variables. The MIP provides many feasible solutions simultaneously with all constrains met, and then chooses the best possible solution (see Breithaupt & Hare, 2007; and Melican, Breithaupt, & Zhang 2010). Despite the published commercial software, such as CASTISEL (Luecht, 1998) and CPLEX 9.1 (ILOG, 2005), MIP can also be conducted in R via the lp_Solve package (Diao & van der Linden, 2011).

2.4 Test Anxiety in Multistage Testing Designs

Test Anxiety is one of the most often-studied topics in educational and psychological measurement (Powers, 2001). It is referred to as an unpleasant emotional state that occurs in an assessment context or evaluative situation (Colwell, 2013) and it is considered as a multidimensional construct (Liebert & Morris, 1967; Sarason, 1984) and

can be defined as a constraining force or psychological influence that can persuade test takers to behave and think differently. Notwithstanding the research over the decades, there is still much uncertainty regarding the role and effect of test anxiety in test taker performance under various testing modes.

2.4.1 Test Anxiety Components in MST

Despite its growing popularity, important technical research issues in MST remain to be addressed for the full potential of MST to be realized. One of the major issues include controlling test anxiety states between test takers by balancing test speededness of preassembled modules. Time pressure is a major component of test anxiety (Orfus, 2008), while it has been shown to increase the rate of performance, existing literature also suggested that the performance in time intensive tests becomes less consistent among students who have high levels of test anxiety. As a result, a significant decline of their performance levels was found compared to their peers who were not affected by test anxiety (e.g., Kellt & Kerau, 1993; Onwuegbuzie, 1995).

Test speededness is a critical factor in operational testing which is often interpreted as time pressure under which test takers have to operate. Given the innate characteristics of the adaptive test, the possibility of differential speededness exists for any testing format in which test takers see individually tailored item sets. In CAT, a fair assessment on a speeded test would ensure that no test taker should be assigned a set of items that takes longer to answer than the item sets delivered to other test takers (Bridgeman & Cline, 2004). Test takers who received a more time-consuming test, that is, a test consisting of a disproportionate number of items that require longer than average to answer, may be at a disadvantage relative to their peers who saw less time-consuming

items. The same logical conclusion holds for MST where test adaptations occur at the module level, testing time pressure difference can be triggered if some modules have items that require more time to solve while the time limit is controlled for all test takers (van der Linden, 2007). As introduced in the earlier section, van der Linden (2007) proposed a detection method that uses item-level residual response times to reveal differential speeded modules in MST.

In addition to test time pressure, numerous studies have shown that using an adaptive test format in high-stake tests is susceptible to the effect of test anxiety (e.g., Pitkin & Vispoel, 2001, Hartig, Moosbrugger, 2009). Compared with the linear test, a decrease in test taking motivation and self-confidence is found to be associated with adaptive testing when test takers are exposed to high difficulty items more quickly than in a linear test (Frey, Hartig, Moosbrugger, 2009; Hausler & Sommer, 2008). According to a written survey on psychological feedback on CAT (Kimura & Nagaoka, 2011), 90% of test takers found the test “difficult” and 60% felt “discouraged” and “unsatisfied” with the experience. To mitigate the adverse psychological influence in CAT, MST was introduced for high-stakes tests by allowing test takers to review items within modules. Nevertheless, administering the test with multiple modules brings up another unneglectable possible stimulus for test anxiety, that is, the change of module difficulty in the course of a testing period.

As introduced earlier, the MST administration is partitioned into several stages. Some test takers could be routed to subsequent modules with noticeable different difficulties after answering items in the previous stage. Despite the extensive literature on

test anxiety in CAT, there has been surprisingly little research conducted for test anxiety manifestation in high-stakes MST due to the module difficulty shift. Regardless of the exact role of test anxiety, there is no doubt that anxiety escalates for a large proportion of test takers in numerous high-stake tests, including the Graduate Record Examinations (GRE) General Test (a standardized measure of verbal, quantitative, and analytical reasoning abilities that is used to facilitate U.S. graduate school admissions). In a survey that assesses test taker perceptions of the revised MST based GRE test (Cline & Powers, 2014), approximately 50% reported feeling at least very anxious after test preparation. Another GRE test taker survey (Powers, 2008) shows that 36% of test takers reported that thoughts of doing poorly during the test interfered with concentration on the test. Once the test takers understand the nature of the MST module selection algorithm, that is, high ability students would be routed to more challenging modules and low correct response rate may lead to easier modules, this survey result raises major concern on differential test anxiety. Because a proportion of test takers who are routed from the moderate module in stage one to the easier module in stage two may realize that they did not perform well in the beginning module. This negative real time feedback would put these test takers in a relatively disadvantaged position as they may not be able to fully concentrate on the test compared to other test takers. The current literature shows a gap in development of test assembly algorithms that take into account psychological differences between the routing pathways that test takers might take during the MST.

2.4.2 Effect of Adaptive Routing on Test Anxiety

To further understand the degree to which employing routing adaptation may heighten the test anxiety in operational MST, two interrelated aspects need to be

addressed: (1) test takers' perceptions of module difficulty change during the test administration and (2) to what extent does the perceived difficulty change elicit anxiety. The discussion on the effect of MST adaptive routing on test anxiety would be more defensible if both aspects can be confirmed.

It has been shown by studies that students are capable of distinguishing test items that are either below or above their abilities. Using a test item difficulty ranking in an undergraduate level test, Morse (2002) found significant correlation between student-ranked five easiest and five most difficult items and the corresponding ten items on the calibrated item difficulty scale. In a more structurally complicated high-stakes MST where items are grouped by difficulty, it is conceivable that how test takers perceive the difficulty of the test items not only depends on their abilities, but the critical time points when routing takes place. That is, the level of test difficulty perceived by test takers at the beginning of the test may be different from that perceived during the test (Hong, 1999), because they might receive various in-test cues or indicators (e.g., estimation of their own test performances). In certain versions of CAT where test takers are allowed to review items, research has focused on the phenomenon that some students used the self-judged current item difficulty to speculate the correctness of the response to the previously administered item (e.g., Vispoel, Clough & Bleiler, 2002). Studies examined the test taker judgement accuracy on item difficulty found that 73% of the difficulty judgements were correct when distinguishing two items whose difficulty parameter differed by 0.50 (Wise, Freeman, Finney & Enders, 1997).

Despite the voluminous research on item difficulty perception in the CAT

scenario, research on the tendency and easiness of distinguishing module difficulty variation in MST is scarce. However, in the process of constructing modules of varied difficulty, two of the following module building characteristics may imply that the difficulty difference at module level is more discernible than the CAT-based item level difficulty difference to test takers. First, testing programs tend to make the modules within each stage more distinct from one another (Zenisky, Hambleton, & Luecht, 2010) to achieve wider coverage across the ability scale. In studies by Xing & Hambleton (2004), the easy and hard modules were centered as much as one full standard deviation apart. Secondly, while the average item difficulties of modules with varied difficulty were set to be apart, the difficulty of items within later stage modules are more clustered than they appear in the starting stage in order to construct a later stage module that is more informative for the provisional test taker ability estimate. Thus, given the difficulty distinction between modules and the item difficulty clusters within modules, it is logical to speculate that test takers are more likely to distinguish difficulty change if they were branched to an easier or harder module in the later stage. Furthermore, unlike the subtle difficulty variation at item level in CAT, MST tends to make difficulty change more obvious by assigning a group of items with similar difficulty in batches. Therefore, it is plausible that test takers would be aware of the module difficulty change when responding to a series of items with lower or higher difficulty in the later stage.

To investigate to what extent the perceived difficulty change elicits anxiety, research has shown that different arrangement of test items (in terms of difficulty) influences in-test anxiety (Tippets & Benson, 1989). With consistent item difficulty arrangement across the test form generally being the least anxiety arousing, other

arrangement alternatives (easy-to-difficult arrangement and difficult-to-easy arrangement) tend to evoke anxiety in varied degree. For MST designs with more than two stages, more complicated module difficulty arrangements are possible, thereby mounting test anxiety across test takers to a further level.

Despite the mixed module difficulty due to the routing pathways, a further possibility that may aggravate test anxiety is that test takers may be mindful of the ordering of tested modules when taking a MST. As reported in GRE test taking preparation study (Cline & Powers, 2014), the most frequent test preparation practice utilized by the vast majority of test takers across ethnicities was to take official GRE practice tests while some of the students devoted more than 100 hours in preparing for the GRE. It can be expected that most of the test takers have an idea of the MST test delivery procedure before entering the testing site. In real time, informed MST test takers can be expected to have a general awareness of the MST delivery procedure, such as its proclivity to route test takers to an easier module after low performance in the first stage, or more difficult module after correctly answering most of the previous items. Consequently, if they encounter an easier module, they may conclude that they have done poorly on the earlier stage, a situation that could increase anxiety for test takers and interfere with their test concentration.

2.4.3 Moderating Test Anxiety with Time Limits

Time pressure is frequently cited to be a potential moderator of the test anxiety-test performance relationship. It has been hypothesized that heightened time pressure may exacerbate the negative influence of anxiety on performance (Chen, 2004). In speeded standardized tests that require one to stay focused for all or most of the testing time, test-

anxious test takers are more challenged than their peers to complete test items within the time limit as they are more likely to be affected by the testing conditions (Zeidner, 1998). Furthermore, under the restrictedly timed condition, test takers may resort to maladaptive strategies to cope with the pressure, slow and overly cautious responding behavior and rapid, careless responses when the time limit approaches, which may both lead to an increased rate of inaccurate response (Geen, 1987).

A large body of literature has shown support of the hypothesis of the moderator role of time pressure in the anxiety-performance relationship. Morris and Liebert (1969) found that test anxiety was negatively related to the scores on a timed intelligence test, whereas the relationship vanished after providing sufficient time. Similarly, using an undergraduate level test that is impossible to finish within the given time limit, Sarason, Mandler and Craighill (1952) found that students with low-anxiety levels performed significantly better when they were told that they were not expected to finish the test compared with their peers who did not receive any instruction. Furthermore, Hill and Eaton (1977) studied the arithmetic problem solving performance between students with varied levels of test anxiety under different time pressure testing situations. Results showed that under a strictly timed condition, high level test anxiety students took longer to answer questions while making more errors compared with their low anxiety counterparts. However, the performance gap was significantly diminished in another comparison group where more test time was provided.

In the context of standard administrations, few studies have focused solely on the impact of the lowered time pressure on test anxiety while the most pervasive research on

time pressure in standardized test is the time accommodation for test takers with disabilities (e.g.,Elliott & McKevitt, 2000; Phillips, 1994; and Ragosta & Wendler, 1992). However, Marquart (2000) investigated not only the impact of extended time on test outcomes, but also on test taking strategies use for students with disabilities and students performing at or above grade level. Results show that extended time changed the way students approached the test and improved students' test taking motivation. Marquart concluded that the lowered time pressure stimulated students to use better strategies, and these in turn reduced test taking anxiety. In addition to the provision of extended time, there is another logical solution for alleviating test anxiety while maintaining the uniform time limit in standard administration. As elaborated in the response time modeling section, items with similar difficulty may require different amount of time to respond (van der Linden, 2006). Therefore, manipulating the item selection algorithm can achieve differential control of test speededness, thereby accommodating students who experience a higher test anxiety level than average.

2.5 Summary

The development of MST is gaining popularity in the educational testing industry as test publishing firms are striving for integrating computer technology with cutting edge measurement tools. This notion should be paired with the increasing research effort into MST with attention centered on specific technical considerations that should not be overlooked.

Developing test assembly strategies that minimize the impact of the measurement instrument attributes that are not relevant to the primary construct of the assessment and

maximize measurement precision is of constant concern for any testing program. Companies can improve the reliability of the assessment by taking into account test taking anxiety that is caused by the employed MST routing procedure. More research needs to be done to determine the best types of MST assembly and how the test can be delivered in the most efficient manner possible. This literature review shows the gap in the understanding of the aggravating effect of current MST designs on test anxiety as well as the development of test configuration methods that can properly handle differential test anxiety. These considerations will go a long way in determining the future of not only MST but the computerized educational test as a whole.

CHAPTER III

METHODOLOGY

3.1 Study Variables, Conditions, and Research Questions

The variables considered in this study can be categorized into three groups: (1) MST design features, (2) item pool characteristics, and (3) speeded response pattern. The MST design features include variables such as the MST panel design (e.g., number of stages, number of difficulty levels per stage, etc.) and test length. Item pool conditions include the distributions of item discrimination and difficulty parameters. In addition, this study also considers the item time intensity parameters since modules with varied time intensity levels need to be constructed to accommodate simulated examinees who exhibit various test anxiety patterns. Lastly, multiple speeded response patterns are included to capture the varied test anxiety states in the course of a testing period. In this study, it is assumed that the time points where examinees perceive the module difficulty change are different. Therefore, the occurrence of speeded response is also varied across examinees. Essentially, this study is the interaction of these three categories that are of interest.

3.1.1 Test Length

Test length is determined by the module length. The majority of the high-quality, high-stake operational MST exams are longer than 30 items. Nevertheless, it is still worthwhile to explore the minimal length of a functional test before the true ability score recovery rate is significantly compromised. Two levels of module lengths are considered (10 and 20 items), consequently, tests under the long test length condition (3 stages) have 60 items. This also mimics the average test length of four Certified Public Accountant

(CPA) examination sections. Additionally, the number of items within each module is also fixed across all conditions.

3.1.2 MST Panel Design

Choices of MST panel design are abundant in the literature and can capture almost any desired administration characteristics by manipulating the number of stages and variety of difficulty levels within stages. As suggested by Amstrong et al. (2004), three modules per stage was sufficient for desirable accuracy of ability estimates for most MSTs. In this study, some common MST panel designs in the research literature are employed. Based on previous studies (e.g., Jodoin et al., 2006; Xing & Hambleton, 2004; Luecht & Nungester, 1998), both two- and three-stage panel designs are considered, as illustrated in the first column of Figure 1, three panel designs: 1-3, 1-2-3, and 1-3-3 are used.

As noted above, in order to accommodate examinees who experience higher real time test anxiety than average, this study proposed an approach to manipulate the item selection in the module generation phase to achieve the real time test-taking anxiety control by adding test speededness constraints to module level. The challenge arises from the need to construct and arrange modules with specified speededness and normal modules within the same panel. For a two-stage design, this issue can be easily solved by replacing the easy module in the second stage with an equally difficult module that is controlled in speededness. Thus, examinees would experience a less time intensive module after being routed from a medium level module. However, in the more complicated 1-2-3 and 1-3-3 design, a single module in the third stage could be assigned

to examinees who previously saw different stage-two modules with varied difficulty levels. Simply replacing a stage-three module may not meet the requirement of speededness control since examinees from stage two need modules with the same difficulty level but different test speededness degrees. Therefore, parallel modules need to be constructed in the third stage to provide modules that are equally difficult but differ in test-taking time pressure.

The proposed panel design is illustrated in Figure 3.1, specifically, blank blocks denote normal modules without any speededness control while striped blocks represent modules with speededness control. As is shown in Figure 3.1, in the second stage of a two-stage design (1-3), the easy module was replaced with an equally difficult module with items chosen for module level speededness control. However, in the third stage of the three-stage MST (1-2-3), a parallel module for module 3M needs to be constructed because the test speededness in stage three for examinees from different stage-two modules (2E-M and 2M-H) is different. Thus, the proposed panel configuration for the 1-2-3 design can administer a less time-consuming module for examinees who are routed from a relatively more difficult stage-two module while maintaining the test speededness degree for other examinees. The proposed parallel module design is the core of this study and both the original and the proposed panel designs are simulated across all simulation conditions so that the estimation performances can be compared comprehensively.

3.1.3 Item Pool Characteristics

This study focuses on the item pool characteristics because item pools are not unlimited resources in operational testing. Therefore, manipulating item pool properties forces the panel assembly process to cope with realities of the item pool. Since all of the

examinee ability scores are sampled from the same normal distribution, $\theta \sim N(\mu = 0, \sigma^2 = 1)$, the potential interaction between examinee ability density, test delivery features, and estimation precision density is indirectly addressed by the varied item pool characteristics. Specifically, this study includes three difficulty distributions with the mean item difficulty, $\mu(b)$. All of the item pools have an average b parameter variation, $\sigma(b)$, of 1.0. Using 1.0 for variance implies moderate variation in the item pool and generates item pools to match the variance of examinees in the population. A constant

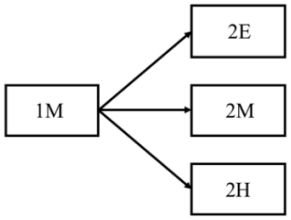
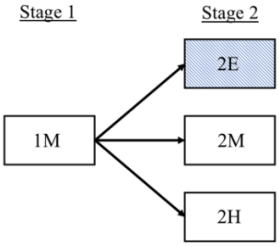
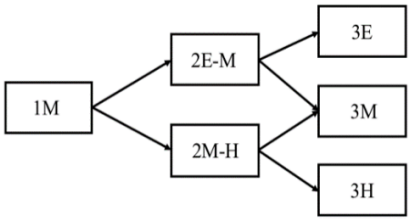
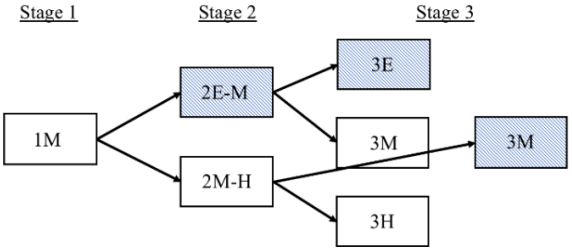
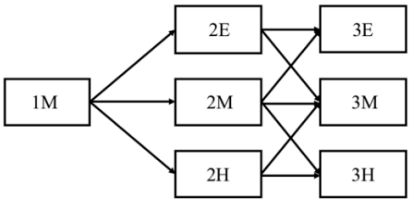
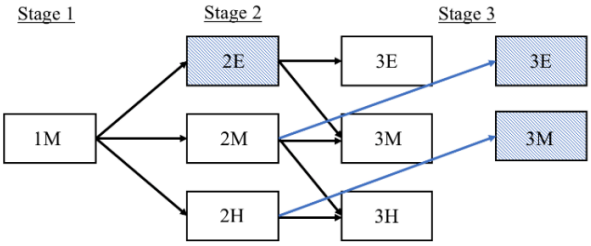
Original panel design	Proposed panel design
• 1-3 Design	
	
• 1-2-3 Design	
	
• 1-3-3 Design	
	

Figure 3. 1 Comparison of module arrangement between original panel design and proposed panel design.

Note. The numbers in each module block denote the stage. E, M and H denotes three difficulty levels (Easy, Medium, Hard). E-M and M-H represent moderately easy module and moderately difficult module, respectively.

level of average item discrimination, $a \sim \text{lognormal}(1, .3)$, and a constant average lower asymptote, $c \sim \text{Uniform}(0, .2)$, , are considered. These values are compatible with a number of large-scale achievement and college entrance examinations and are chosen to mimic an operational computerized test pool (Leung, Chang & Hau, 2005).

Besides item difficulty and discrimination, time intensity parameters are also simulated so that the module speededness level can be controlled in the panel configuration phase. To study the impact of dependency between item difficulty and time intensity, for each item, the difficulty parameter, b, and time intensity, β, are drawn from a multivariate normal distribution with covariance matrix of Σ. The diagonal elements of Σ are 1 while the off-diagonal values are varied with two conditions: 0 and .33 as the correlation between the two parameters. Particularly, since the positive correlation between the item difficulty and the item time intensity is often found across items (e.g., van der Linden et al., 1999; Klein Entink et al., 2009), the same pattern can be expected within a specific item: If an item is difficult for a given test taker, it will more than likely require longer to complete. In an empirical analysis of the relationship between the item difficulty and the item time intensity using the data from a computerized Certified Public Accountant Examination, van der Linden (2007) found the more difficult items tended to be more time intensive while the correlation between these two parameters were .30. Therefore, .33 was chosen as one of simulation conditions to describe the relation

between item difficulty and item time intensity. Furthermore, controlling the item parameter distributions allows for the examination of proposed MST panels under different item supply conditions.

3.1.4 Speeded Response Pattern

While test anxiety is a fairly stable attribute, test anxiety state tends to be volatile over time (Hong 1999). To realistically capture the response pattern of examinees who experience test anxiety, this study considers three speeded response patterns during an exam. According to the study conducted by Ping (2008), test anxiety pattern fluctuations often manifest as two trends: (1) examinees experience the peak of the anxiety symptom at the beginning of the test. As they proceed with their tests, both physical and psychological symptoms gradually taper down; and (2) For examinees with higher anxiety trait, the declined anxiety state is usually followed by a resurgence of anxiety symptoms.

As mentioned in the previous section, a heightened anxiety state may lead to inconsistent or lower performance during a test, which in turn forces examinees to use longer time on each item. Therefore, in original panel designs where speededness control was not in place, examinees who routed from a relatively harder module to an easier module may encounter test speededness and its degree may vary in different levels. This study simulates the impact of varied anxiety states on examinees' responses by allowing the speededness effect to emerge at different test locations for a specified group of examinees. That is, a certain proportion of these examinees' responses are generated to be speeded responses. In other words, the probability of correctly answering the given

items would be lower when responses are speeded compared with their unspeeded counterparts.

More specifically, three speeded response patterns are considered in the study. Firstly, examinee's responses to 30% of the items in the module are simulated to be speeded in order to capture a moderate level of test speededness which is caused by the perception of module difficulty change. For the second speeded response pattern where symptom resurgence appears, both the first and the last third of the responses in the given module were simulated to be speeded. Lastly, the third speeded response pattern is simulated to represent the scenario that the adverse effect of test speededness not only has an influence on examinee performance in the current module, but also lower their performances in the following module. Therefore, the speeded responses are generated for 30% of the items in the current module plus 30% of the items in the following module regardless of routing direction. The third scenario of speeded response pattern represents the aftermath of negative direction routing for those examinees with high test anxiety traits and their responses would still be speeded even though they were routed to a higher-level module. It also needs to be noted that the third speeded response pattern only applies to the panel configurations with three stages. Specifically, it is designed to capture the performance of students who were routed to a relatively easier module in stage two and then routed to a lower or a medium level module in the third stage. Details of speededness effect modeling and speeded response generation are provided further on in the later section (see **3.3 Statistical Methods**).

3.1.5 Simulation Conditions

The complete set of crossed design conditions is summarized in Table 3.1. As is shown in the table, a total of 108 conditions are addressed by the study. Based on common practice of MST simulation, each of these conditions are replicated 10 times with the same item pool determined by the item pool characteristics.

3.1.6 Research Questions

Based on the overall purpose of the study and the consideration of study variables introduced above. Two research questions were proposed:

1. To what extent does the proposed MST design recover the simulated proficiency scores under a variety of panel configuration conditions?
2. How do various item pool characteristics affect the psychometric quality of the proposed MST design?

Table 3. 1 Summary of Study Design Conditions

MST Designs		Item Pool Characteristics		Speeded Response Pattern
Module Length	Panel Design	Average Item Difficulty	Correlation Between Difficulty and Time Intensity	Speededness Effect Occurrence
10	1-3	$\mu_b = -1$	$\rho_{\beta b} = 0$	Effect occurs at the beginning of the module for 1/3rd of the module length
20	1-2-3	$\mu_b = 0$	$\rho_{\beta b} = .33$	Effect occurs at both the beginning and the end of the module for 2/3rd of the module length
	1-3-3	$\mu_b = 1$		Effect occurs at both the current module and the following module for 2/3rd of the module length in total

All research questions will be addressed using a simulation approach that will include a fully crossed design, 10 parallel panels will be assembled of each condition and using the same examinees under each of the three MST panel design pairs (original design and the proposed design). By holding the examinees constant across panels design, it helps to directly control for sampling errors in the methodological comparisons.

3.2 Data Generation and Simulation Procedure

Across all 108 distinct conditions, each simulation is conducted in three steps: (1) Item generation and calibration; (2) panel assembly; and (3) MST administration simulation. Details of each step is described below. All of the psychometric procedures introduced below are conducted in R (R Core Team, 2014).

3.2.1 Step One: Item generation and calibration

For each of the simulation conditions, an item pool is specifically created with respect to the panel design. The size of the item pool is set as 1.5 times the number of items needed for a typical paper-pencil fixed length test (Wang et al., 2012). The prespecified item parameters are used to generate a response data set for calibration. Specifically, 500 simulated examinees ($\theta \sim N(0, 1)$) are assigned to each of the 15 sets of linear form test items, while each set matches the corresponding panel test length in the given condition. For instance, in the 1-2-3 panel design with 10 items per module, a typical simulated examinee would be assigned a 30-item linear form test to mimic the test administration of three modules from each of the three stages. To allow for 10 parallel panels of each condition set, the item pool size will be $60 \times 10 \times 1.5 = 900$, where 60 is the number of items needed to assemble a 6-module panel given the 1-2-3 panel

configuration. This item generation and calibration approach simulates a realistic procedure to calibrate an item pool based on field testing.

In addition, in order to simulate item pools with realistic item pool sizes and create challenging administration conditions for the proposed MST panel design, item pools that require more than 1050 items will be limited at the 1050 item supply. Thus, all conditions with module length equal to 20 will not use an ideal item pool based on the ideal item pool size calculation above. Table 3.2 presents the summary of item pool size for various panel configurations.

Table 3. 2 Summary of Simulated Item Pool Size by Panel Design and Module Length

Panel Design	Module Length	Ideal Pool Size	Simulated Pool Size
1-3	10	600	600
1-2-3	10	900	900
1-3-3	10	1050	1050
1-3	20	1200	1050
1-2-3	20	1800	1050
1-3-3	20	2100	1050

3.2.2 Step Two: Panel Assembly

Using the calibrated item statistics from step 1, ten parallel panels are assembled from the generated item pool. The maximum of item overlap frequency across panels is fixed at half of the panel count, which equals 5. The panel assembly can be broken up into two parts: Test Information Function (TIF) targets selection and Automated Test Assembly (ATA). In this study, the Bottom-Up assembly strategy was employed for panel assembly. This strategy is capable of constructing parallel panels simultaneously.

As suggested by Luecht (2006), separate information targets are required for each module. The approach for specifying TIF targets is described below:

1. Based on the item statistics of the generated item pool, the location and the value of maximum information are computed. For example, in an item pool with a generated average difficulty of 0, the location of maximum information is likely to locate around 0.

2. Items in the pool are sampled into N bins where N = the number of stages (2 or 3). The number of items sampled to each bin is proportionally equal to the sum of the module sizes per stage. Moreover, the location of maximum information of each bin should be consistent with calculated location in step 1. Using the previous 1-2-3 design sample (item pool size = 900) as an example, three bins are generated corresponding to three stages. 1/6th of the item pool (150) would be sampled in bin #1, 1/3rd of the items in the pool (300) would be sampled in bin #2 and the remaining half of the items would go to bin #3.

3. Items in each bin are sorted by location of maximum information and items are split into D “piles” where D = the number of difficulty levels per stage. In the 1-2-3 design, no split is needed for bin #1 since it only has one difficulty level. Items in bin #2 would be sorted and split into a moderately easy pile and a moderately hard pile. Lastly, three piles are needed for bin #3 to represent easy, medium and hard difficulties.

4. The average item information across all items within each pile is computed at 41 quadrature points on the ability scale (θ from -4 to 4, by 0.2). Then, the average item information across each pile becomes the TIF target for ATA procedure.

Once the target TIFs are determined, this study uses Mixed-Integer Programing (MIP) as the ATA procedure to simultaneously build multiple panels. Specifically, the R Package *lpSolveAPI* (Konis, 2009) was used to solve the multiple test form assembly problem for an optimal solution. This algorithm can also handle any number of constraints. As mentioned in the MST panel design section, this study proposed panel designs with the speededness-control modules. Specifically, some of the items in these modules need to be less time intensive while remaining equal in difficulty. To realize this module design in test assembly, the MIP procedure can be configured to construct the module with specified proportion of the items are selected from the lower half of the time intensity scale. Details of the MIP procedure is introduced in a later section (see **3.3 Statistical Methods**).

3.2.3 Step Three: MST administration

The last step uses the pre-assemble panels from step 2, for each of the 10 parallel panels, 2000 examinees will be simulated. The estimated module parameters and item statistics from previous steps are used for routing scoring while the response data are generated using the true item level statistics. This allows estimation error to play a role in scoring accuracy, in addition to the errors attributable to the routing and scoring mechanism. Specific routing and scoring rules are introduced below.

For each simulated examinee, the response generation process can be described as follows, given a sampled proficiency score, θ. For each person-by-item administration, the "true" probability of correct response, P_{ij} ,can be computed. Then, a random number, π_{ij}, is generated from a uniform distribution within the interval $0 < \pi_{ij} < 1$. If $P_{ij} > \pi_{ij}$, we set the scored item response to 1, otherwise 0. Besides the response generation under 3PL model, the examinees' responses under speeded response pattern is generated via the Multi-class Mixture Rasch Model (MMRM). The full-pool response generation is used again for data management convenience. The response generation process for MMRM is introduced in a later section (see **3.3 Statistical Methods**).

The Approximate Maximum Information (AMI) method is employed to determine the routing point. AMI method defines the routing points as the intersection point of the test information curves of the previously administered and current module (Luecht, Brumfield & Breithaupt, 2006). Figure 3.2 demonstrates an AMI method for a 1-3-3 design. Two routing points, θ_L and θ_U, are determined at stage 2. θ_L denotes the test information curve intersection between routing pathways 1M+2E and 1M+2M (E and M represents an easy module and a moderate difficulty module, respectively), while the θ_U denotes the intersection between routing pathways 1M+2M and 1M+2H. Examinees with estimated proficiency scores lower than θ_L are assigned to 2E and their peers with scores higher than θ_L are routed to 2H. The rest of the examinees are routed to 2M. Given the specified routing points, the present study scores the simulated examinee with the IRT-based expected a posteriori (EAP) method.

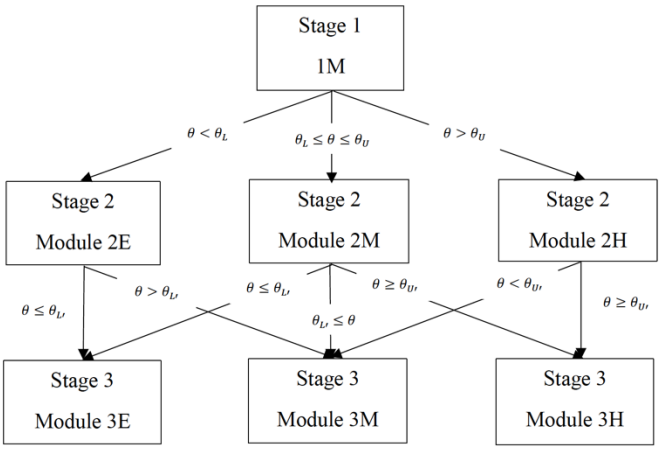


Figure 3. 2 Example of AMI procedure.

Note. The numbers in each module block denote the stage. E, M and H denotes three difficulty levels (Easy, Medium, Hard).

Lastly, the simulation process for each of the 108 conditions can be summarized as follows:

1. True item parameters based on 3PL model are generated with the total number of items to match the specified item pool sizes.
2. Linear form tests are assembled using the randomly selected items from the generated item pool. The length of the linear form test matches with the sum of total number of items used in a specific MST design.
3. 500 examinees are simulated to respond to each of the linear forms. Item parameters are estimated using the 3PL model.
4. Given the estimated item level parameters, ATA procedures are employed to assemble 10 parallel panels using the identified Test Information Function targets.
5. 2000 simulated examinees are randomly assigned to each of the 10 parallel panels.

6. Responses for unspeeded items are generated with the 3PL model and MMRM model is employed for generating speeded responses.
7. AMI method is performed for routing examinees in the assigned panel.
8. The proficiency score from the final stage is estimated and recorded.

3.3 Statistical Methods

3.3.1 Speeded Response Generation

To simulate responses under the speeded module, this study uses the Multi-Class Mixture Rasch Model (MMRM; Morch, Bolt, & Wollack, 2005) to generate responses when the speededness effect emerged during the administration. Specifically, the MMRM is used to generate speeded responses in two module-level locations: the beginning of the module (1/3rd of the module length) and both the beginning and the end of the module (2/3rd of the module length). The MMRM model is chosen over the two-class Mixture Rasch Model (MRM; Bolt et al., 2002) and the HYBRID model (Yamamoto, 1997) due to two reasons: (1) the MRM is ruled out because the present study includes more than two item latent class profiles while MRM can only capture two-profiles. (2) the HYBRID model is not considered because it treats speeded responses as random, however, the effect of speededness is more often manifested as forcing examinees to spend longer times on items.

Essentially, the MMRM assumes that examinees belonging to the common latent class profile experience the same level of test speededness throughout the course of administration. By assuming the effect of speededness can be reproduced by altering item difficulty parameter of an item, the MMRM allows examinee ability to be relevant to

item responses that are speeded. Under the MMRM, the probably that an examinee *j* correctly answers an item *I* is expressed as

$$P(u_{ij} = 1 | \theta_{gj}, b_{gj}, g) = \exp(\theta_{gj} - b_{gj}) / [1 + \exp(\theta_{gj} - b_{ig})], \quad (3.1)$$

where

u_{ij} is the response outcome of examinee *j* to item *i*,

θ_{gj} is the latent ability parameter of examinee *j* in class *g* {*g*=1, 2, …, *k* latent classes},

and

b_{ig} is the difficulty of parameter of item *I* in class *g*.

The key difference between the MMRM and the Rasch model is the subscript *g*, which indexes item latent classes. Under the MMRM, multiple Rasch person ability estimates can coexist by specifying various latent classes, *g*. In the present study, item locations are required to differ across class at two test locations: the beginning of the test (1/3rd of the module length) and both the beginning and the end of the test (2/3rd of the module length). Table 3.3 illustrates the MMRM application of the latent class profile in a 20-item module with 3 latent classes. For all examinees, responses to items in the middle part of the module are controlled as unspeeded responses while the responses to other items are allowed to be speeded across profiles.

Table 3. 3 Latent class profile for an MMRM speededness model with 20 items and 3 latent classes.

	Items											
Latent Class	1	2	…	6	7	8	…	14	15	16	…	20
1	U	U	…	U	U	U	…	U	U	U	…	U
2	S	S	…	S	U	U	…	U	U	U	…	U
3	S	S	…	S	U	U	…	U	S	S	…	S

Note. U = Unspeeded item; S = Speeded item

To implement the MMRM to generate speeded responses, items with speeded latent profile need to be coded with multiple difficulties to fit various examinee response patterns. However, the implementation of the MMRM in the present study is simplified due to the simulation design. Since the full-item pool response generation is adopted for speeded response generation, the difficulty parameters of the entire pool can be uniformly altered to generate speeded responses.

The magnitude of item difficulty change, $b_{1j} - b_{2j}$, is determined based on the simulation study conducted by Morch et al. (2005), where they employed a speededness simulator (Wollack & Cohen, 2004) to estimate the item difficulty change in various speeded response conditions. The speededness simulator assumed that speededness emerges at different points for different examinees and is manifested by a substandard performance as time runs out. From the simulation results, the range of item difficulty change due to speeded response is between 0.07 to 1.8 with the average of .51. Therefore, to mimic the speededness simulator mechanism, item difficulty changes between speeded and unspeeded response in the present study is randomly sampled from the distribution, $b_{1j} - b_{2j} \sim N(.5, .2)$.

3.3.2 MIP Method

In order to generate multiple parallel forms in MST that share a set of psychometric characteristics, the Mixed Integer Programing (MIP) method was used in this study. Particularly, the R Package *lpSolveAPI* (Konis, 2009) was used to solve the multiple test form assembly problem for an optimal solution. This section introduces the modeling principles of the MIP method and uses one of the panel configurations as an example to illustrate the application of MIP.

Building multiple alternate test forms is commonly required in educational testing, its goal is to group a set of items from the item bank to meet a variety of test form constraints (e.g., test length, difficulty, content). To automate this task, various automated test generation methods were developed, and they treated the test assembly problems as optimization problems with the goal to construct the best test form from a set of alternatives while meeting various constraints. MIP method is one of the methods using mathematical programing to solve the test assembly problem.

In general, the MIP method uses a decision variable as an unknown decision that are to be optimized. In the case of automated test assembly, the binary decision variable is defined as whether item X_i is assigned to form *f*. Then, an objective function is defined to assess the quality of the solution and the program will either minimize or maximize the value of the objective function. To build multiple forms with equally high test information values, for instance, the outcome of the decision variable is evaluated so that the overall test information of the assemble forms are maximized. Finally, the decision that must be made is subject to a series of requirements, such as test length, content

bounds, and mutually exclusive items. Each of the constrains calls for a linear function that depicts either an equality or inequality relationship between the decision variable and a scalar value.

An example of 1-3 panel design assembly (Diao & van der Linden, 2011) is used in this section to illustrate the procedure of ATA via the MIP Method. Specifically, the Test Information Function (TIF) for the stage-one routing test was required to meet a relatively uniform target at $\theta_k = -1, 0, 1$. The second stage TIF was specified to have single peak at $\theta_k = -1, 0, 1$ respectively to represent the modules of low, moderate and high difficulties. The test length was fixed at 20 for modules in both stages. Item overlap between modules was not allowed.

The decision variable x_i are defined as

$$x_i \begin{cases} 1 & \text{if item } i \text{ is assgined,} \\ 0 & \text{otherwise.} \end{cases} \tag{3.2}$$

The $I_i(\theta_k)$ is used to denote the information function value at θ_k for item *I.* To assemble the first stage modules with a set of uniform targets, the maximum principle was used for objective function to bring the minimum value for the TIFs at the three specified θ_k (-1, 0, and 1) as high as possible. In addition, the tolerance parameter, δ, was set equal to 0.5 to set an extra requirement to ensure a uniform distirbution. Formally, the MIP model used for stage one assembly was

$$\max y \tag{3.3}$$

subject to

$$\sum_{i=1}^{I} I_i(\theta_k)x_i \geq y, \text{ for all } k, \tag{3.4}$$

$$\sum_{i=1}^{I} I_i(\theta_k)x_i \leq y + \delta, \text{ for all } k, \tag{3.5}$$

$$\sum_{i=1}^{I} x_i = 20, \tag{3.6}$$

$$x_i \in \{0, 1\}, \tag{3.7}$$

$$y \geq 0, \tag{3.8}$$

The combination of objective function 3.3 and constraint function 3.4 uses maximum principle to obtain a uniform distribution over three specified θ points with the maximum test information. Particularly, objective functions 3.4 and 3.5 ensure the sum of item information of the selected items at three θ_ks (k = -1, 0, and 1) are within the 0.5 range between y and $y + 0.5$. Thus, by maximizing y, the TIFs at these three thetas are relatively similar while reaching the highest maximum. In addition, the module length was controlled by equation 3.6.

For three second stage modules, they were designed to produce a peaked TIF at three different locations. The model is defined as

$$\max y \tag{3.9}$$

Subject to

$$\sum_{i=1}^{I} I_i(\theta_k)x_{if} \geq y, \text{ for all } k \text{ and all three forms,} \tag{3.10}$$

$$\sum_{i=1}^{I} I_i(\theta_k)x_{if} \leq y + \delta, \text{ for all } k \text{ and all three forms,} \tag{3.11}$$

$$\sum_{f=1}^{3} x_{if} \leq 1, \text{for all } i, \tag{3.12}$$

$$\sum_{i=1}^{I} x_{if} = 20, \text{for all three forms,} \tag{3.13}$$

$$x_i \in \{0, 1\}, \tag{3.14}$$

$$y \geq 0, \tag{3.15}$$

Compared with the model for stage 1, the main difference is the subscript *f*, which denotes three modules with different difficulty requirements in stage 2. Moreover, the tolerance parameter δ, was set to be 0.2 in stage 2. Equation 3.12 ensures no item overlap occurs between three modules by forcing the selection count of a typical item across three forms to be equal or less than 1.

3.4 Data Analysis

Once the estimated proficiency score, $\hat{\theta}$, is calculated for each examinee, residual statistics will be calculated within each of 108 distinct conditions. For each $\hat{\theta}$, the residual for the *j*th examinee can be calculated as $e_j = \hat{\theta}_j - \theta_j$. The mean bias is calculated as

$$\mu_e = E_j = \frac{\sum_{j=1}^{N} e_j}{N}. \tag{3.16}$$

Bias often serves as a measure for detecting Differential Item Functioning. However, since there is only one group in the study, bias then implies the extent to which all examinees with similar abilities perform at the same level. In other words, does the choice of MST panel design cause groups of examinees with the same proficiency score to be misestimated?

Another outcome variable is also related to the residual error. Root Mean Square Error (RMSE) is defined as

$$RMSE = \sqrt{\frac{\sum_{j=1}^{N} e_j^2}{N}}. \quad (3.17)$$

The expectation of RMSE in a distribution with no error is 0. In this study, the RMSE will be used to compare the true theta to the estimated theta and to evaluate how well the ATA generated Module Information Functions (MIF) matching the target MIFs.

The Mean Absolute Error (MAE) is also reported. MAE measures the average magnitude of the errors in a set of predictions, without considering their direction. Formally, it is expressed as

$$MAE = \frac{\sum_{j=1}^{n} |e_j|}{n}. \quad (3.18)$$

The MAE and the RMSE can be used together to diagnose the variation in the errors in a set of forecasts. The RMSE will always be larger or equal to the MAE; the greater difference between them, the greater the variance in the individual errors in the sample. If the RMSE=MAE, then all the errors are of the same magnitude. Both the MAE and RMSE can range from 0 to ∞. They are negatively-oriented scores: lower values are better.

Beside the aforementioned evaluation indices, several item pool usage indices will also be calculated in this study to examine the different MST configuration performances via the perspective of item pool use efficiency. Specifically, the item pool usage indices include a) percentage of item pool used; b) max item exposure rate; c) mean item exposure rate; and d) standard deviation of item exposure. The item exposure

rate was computed by dividing the number of times the item is administered by the number of examinees (e.g., 2000).

3.5 Review of Research Questions

1. To what extent does the proposed MST design recover the simulated proficiency scores under a variety of panel configuration conditions?

The three outcome statistics mentioned earlier will be computed and graphics will be plotted to demonstrate the overall effect of the proposed MST panel design under a variety of panel configuration conditions. The results from this question will directly address the research gap in the literature on the feasibility of using the proposed panel design to control response speededness.

2.How do various item pool characteristics affect the psychometric quality of the proposed MST design?

Graphics will be developed to demonstrate under what item pool design conditions do the MST panel configurations tend to work best from an estimation accuracy perspective. The graphics from this question will also examine potential problematic item pool characteristics that can arise as a direct reason for substandard model performance. In addition, conditional residuals will be calculated, and plots of residuals will be produced to examine the item pool impact under various panel design conditions.

CHAPTER IV

RESULTS

The results from the Automated Test Assembly (ATA) and simulation study are presented in this section. The evaluation of the ATA results includes how well the specified ATA-generated module information functions (MIFs) fit to the targets across varied test configurations and item pool characteristics. Then, the results from the MST simulation are presented.

4.1 MIFs Generation

The results from the module information functions (MIFs) is discussed in this section and it shows how accurately the generated MIFs from the ATA procedures fit the pre-specified test form information target. Figures 4.1 and 4.2 present an example of one of the comparison plots between generated MIF and target in a two stage (1-3) panel configuration. The plot to the left in the figure denotes the pre-specified information target, while the plot to the right shows the actual information of the assembled modules from each of the 10 generated parallel panels given the MST configuration. The ability levels defined by the generated item pool are plotted on the x axis and the y axis represents the achieved information. The y axis scale goes from 0 to the maximum information obtained.

If the actual information plots resemble the corresponding target plot, it means the generated MIFs fit well with the target. Inspection of Figures 4.1 and 4.2 shows that the actual ATA-generated MIFs approximate to target not only from the shape of the information curve but also the achieved information level. As seen in Figure 4.1,

illustrating the simulation condition of L1C1D2R1 (Module Length = 10, Panel Configuration = 1-3, Average Pool Difficulty = 0, and Correlation Between Difficulty and Item Time Intensity is 0), the fitting was most varied across panels between the ability levels of -1 and 1. It is due to the ATA setting of limiting the item usage frequency. In conditions that sub-optimal item pool sizes were employed, such as 1-2-3 design with module length of 20, the information level can still be retained as the max item usage frequency was fixed at the half of the parallel panel count.

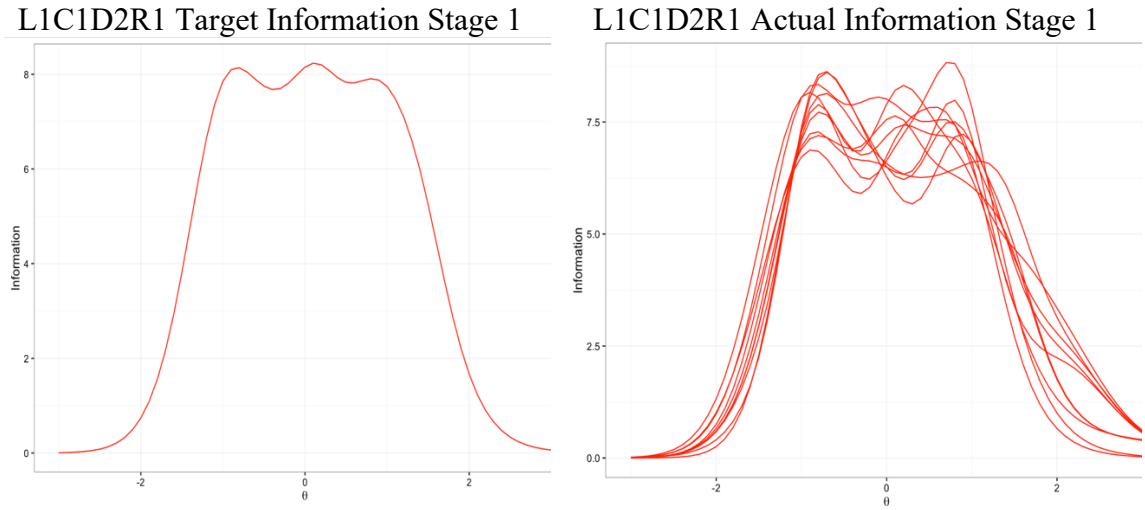


Figure 4. 1 Example of Stage 1 Module MIF Generation from ATA Procedure in a Two-Stage Configuration.

*Note:*The plot to the left in the figure denotes the pre-specified information target, while the plot to the right shows the actual information of the assembled 10 modules.

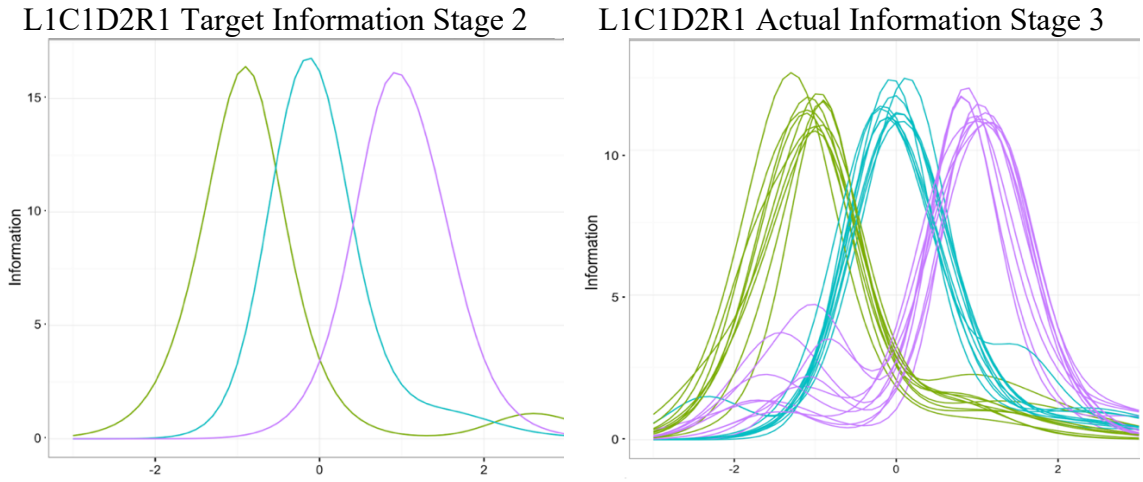


Figure 4. 2 Example of Stage 2 Module MIF Generation from ATA procedure in a Two-Stage Configuration.

Note: Green indicates the low difficulty module, blue indicates moderate difficulty module, and purple indicate the most difficult module. The plot to the left in the figure denotes the pre-specified information target, the plot to the right shows the actual test information of the assembled 10 modules.

The overall RMSEs of the actual generated MIFs to the target MIFs by panel configuration are presented in Tables 4.1 to 4.3 (see Apendix). The simulation conditions were expressed by the abbreviations of each of the four study variables. Specifically, three panel configurations (1-3, 1-2-3, 1-3-3) were denoted by C1, C2 and C3; two module lengths (10, 20) were denoted by L1 and L2; three overall item pool difficulties (-1, 0, 1) were represented by D1, D2 and D3; two correlations (0, 0.33) between difficulty and time intensity were represented by R1 and R2. These RMSE indices are summed over the entire range of the ability scale. Across the scale of ability, considerable amount of misfit towards both tails of the distribution was observed while significantly less misfit was found around the peak of the distribution. This could be an artifact of the item parameter generation since less items were generated at both tails of the logit scale given the parameter generation approach. In general, the module produces the biggest difference in fit. The modules in the 1-3 configuration tend to fit relatively better than the

modules from either the 1-2-3 or the 1-3-3, while the latter two configurations fit equally well in terms of the RMSE. Overall, a simpler panel design with a 10-item module length had a better chance of fitting compared with other conditions.

4.2 Impact of Speeded Response on Routing

In the present MST administration, the only way that the ability estimates of two equally abled examinees differ is if their test routings change as a function of the in-test speeded response patterns. That is, when the anxiety-caused test speededness is presented to an examinee, the examinee's score may be underestimated. Therefore, he or she is more likely to be routed to a lower level module in the following stage while an equally abled examinee who doesn't experience test speededness can be routed to a module that matches better with the examinee's true ability. Figure 4.3 illustrates the effect of the anxiety-caused speeded response patterns on the administration routing with the 1-2-3 panel configuration as an example. The proportion of examinees who routed to each of the four possible paths are plotted by speeded response pattern and average item pool difficulty. The complete path-taking proportions across all conditions is presented in Tables 4.8 - 4.10 (see Apendix).

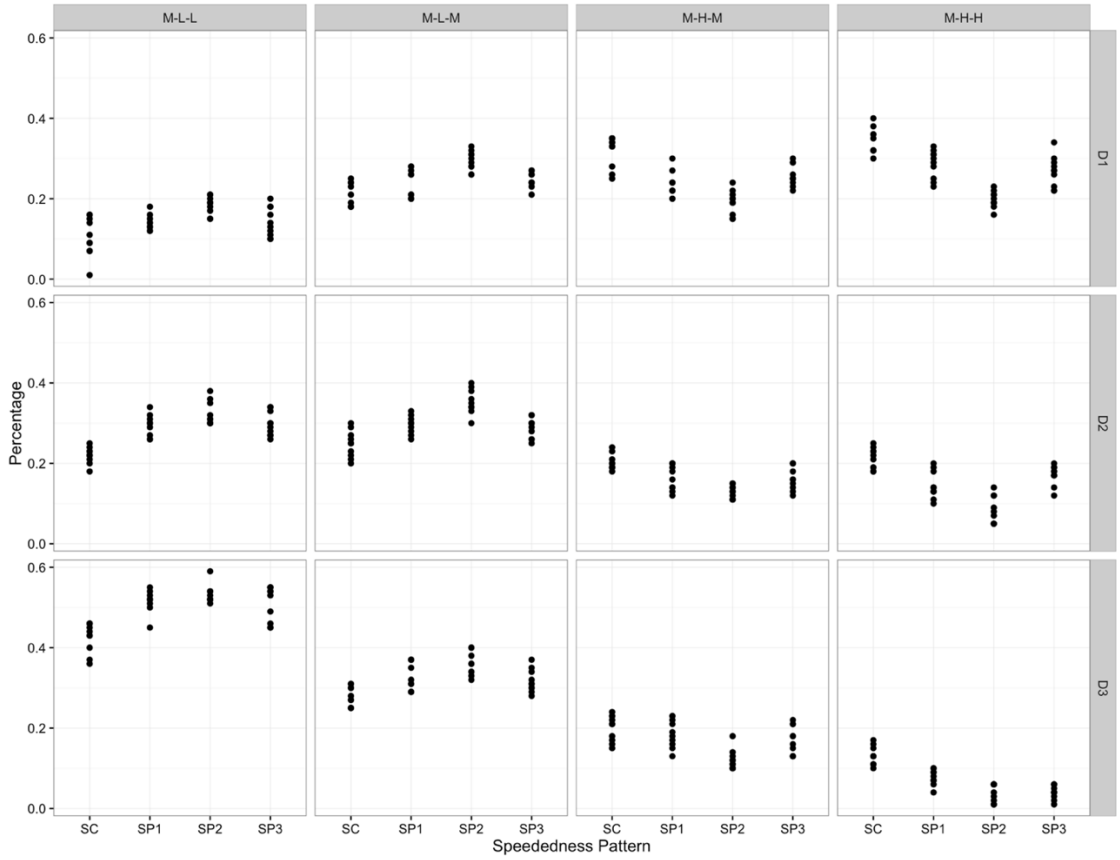


Figure 4. 3 Proportion of routing path taken by simulated examinees across speeded response patterns and item pool difficulties in a 1-2-3 design.

Note: The names of the four speeded response patterns have been abbreviated: SC = Speededness Control, SP1, SP2, and SP3 denote three speeded response patterns. Three overall item pool difficulty (-1, 0, 1) are denoted as D1, D2 and D3.

Overall, the speededness control (SC) group seems to be working appropriately as it routes roughly 25% of examinees to each one of the four paths when the item pool difficulty is moderate (D2). As the center of item pool difficulty is shifted to -1.0 (D1), the speededness control group routes relatively more examinees to M-H-H path while less examinees are routed to modules with more items of lower difficulties. The opposite routing pattern is also observed when the mean of item pool difficulty was lifted to 1.0 (D3), that is, examinees are routed to the M-L-L path more often than when the average

difficulty of the item pool is centered at 0. There are cases where the proportion of examinees taking that route is almost 60%.

Compared with the speededness control group, three speeded response patterns (SP1, SP2, and SP3) present noticeable differences of group level path-taking proportions across paths and average item pool difficulties from the speededness control group. Examinees' scores are underestimated across item pool difficulty levels because of the effect of speededness. The second row of Figure 4.3 displays the path-taking proportion when the average pool difficulty is centered at 0. More examinees were routed to either the lower level path (M-L-L) or the path which includes lower level module (M-L-M) while examinees tended to lose the opportunity to be routed to paths with higher level module (M-H-M and M-H-H). The first and the third row of Figure 4.3 further highlight the impact of speededness by shifting the mean difficulty of item pool off-center. A smaller proportion of examinees was routed to the lower level path (M-L-L) when the overall test difficulty is -1 and only about 10% of the examinees took the higher level path (M-H-H) when the overall test difficulty was fixed at 1.

To compare the path taking proportion within three speeded response patterns (SP1, SP2 and SP3), it can be seen from the Figure 4.3 that the SP2 consistently shown more deviation from the speededness control group than SP1 and SP3, which means that the within stage speeded response proportion plays a stronger role in assigning examinees to a suboptimal module . Moreover, the path taking proportions across simulation conditions were nearly identical between SP1 and SP3. Taken together, these results suggest that there is an association between the percentage of speeded responses and the

group level path-taking proportion. That is, as the proportion of speeded responses increase, examinees are more likely to take paths that consist of lower level modules.

4.3 Examinee Ability Recovery

Ability score recovery is presented in this section to demonstrate to what extent the proposed models provide accurate estimates about the examinee's ability level. The precision of score estimation is examined across all conditions: module length, panel configuration, mean item pool difficulty, correlation between item difficulty and item time intensity, and speeded response pattern. First, the multi-factor ANOVA results using three evaluation indices (bias, RMSE, MAE) will be examined. Next, figures of bias, RMSE, and MAE conditioned on each of 31 quadrature points across the theta scale will be shown. The detailed information of the evaluation indices for all simulation conditions was summarized in Tables 4.11 to 4.18 (see Apendix). Each of these figures includes three speeded response patterns, and each plot also contains the results of the "reference group" which is the MST administration with the proposed speededness control panel configuration. The colors are consistent across all of the plots with a color per pattern: red represents the speededness control group, green indicates the first speeded response pattern (SP1: 30% of the responses in the current module are speeded), blue and purple indicate the second and the third speeded response patterns respectively (SP2: 60% of the responses in the current module are speeded; SP3: 30% of the responses in both current and the following module are speeded).

Table 4.4 shows the correlations between the true examinee's ability score and the estimated ability score across all simulation conditions. It was found that the correlation of the true ability and the estimated ability did not vary significantly within the three

speeded response patterns when the other three conditions were controlled. Therefore, the correlations for three speeded response patterns was summarized as the averaged correlation when compared with the correlations in the speededness control condition.

Table 4.4 Correlation between the True Ability Score and the Estimated Ability Score

Simulation Conditions				Correlation
MODLEN	CONFIG	COR	SP	
10	1-3	0	SP	0.893
		0.33		0.905
		0	SC	0.911
		0.33		0.909
	1-2-3	0	SP	0.886
		0.33		0.894
		0	SC	0.928
		0.33		0.932
	1-3-3	0	SP	0.873
		0.33		0.881
		0	SC	0.930
		0.33		0.935
20	1-3	0	SP	0.871
		0.33		0.875
		0	SC	0.929
		0.33		0.930
	1-2-3	0	SP	0.853
		0.33		0.849
		0	SC	0.948
		0.33		0.941
	1-3-3	0	SP	0.857
		0.33		0.855
		0	SC	0.952
		0.33		0.949

Note: The names of four simulation conditions have been simplified: MODLEN = Module Length; CONFIG = Panel Configuration; COR = Correlation between Item Difficulty and Item Time Intensity; and SP = Speeded Response Pattern.

In general, when the speeded response pattern was presented, longer test length didn't improve the correlation. Contrastingly, it lowered the correlation slightly when the module length or the number of stages increased. In addition, the correlation between

item difficulty and item time intensity did not seem to affect the ability recovery. However, for speededness control condition, the estimation precision improved as the test length was extended. Similarly, the correlation between item difficulty and item time intensity did not play a role in ability recovery.

4.3.1 Multi-Factor ANOVA Results

Tables 4.5 to 4.7 present the multiple factor ANOVA results using three evaluation indices (bias, RMSE, and MAE) as the criteria. In addition, the effect size, η^2, is reported as a measure of the degree of association between effect (e.g., a main effect, an interaction) and the dependent variable. From the *F* test results, the main effect of the module length was marginally significant with the all three criteria, as was the panel configuration and the speeded response pattern. The correlation between item difficulty and item time intensity, however, did not yield a significant *F* ratio when using bias and MAE as the criteria, indicating that the mean bias of aggregated ability score estimates were not significantly different when the correlation between difficulty and time intensity changed from 0 to 0.33. Regarding the two-way interaction terms, all interaction terms including the module length (MODLEN) were significant, indicating that the effect of module length interacts with panel configuration (CONFIG), average item pool difficulty (DIFF), the correlation between item difficulty and time intensity (COR), and the speeded response pattern (SP).

Besides the *F* test results, the effect size (eta squared) is reported to represent the proportion of variance in the dependent variable (evaluative indices) that is attributable to each effect. It is found that some significant effects (e.g., CONFIG, MODLEN:COR)

don't account for much variance. Using 3% of variance accounted for as the criterion, some significant terms can be ignored. However, the module length and the speeded response pattern are consistently significant. Specifically, the module length accounted for approximately 10% of the total variability in three evaluative indices. Speeded response pattern was found to account for much larger proportion of variability across three evaluative indices compared to module length.

Table 4.5 Fixed-Effect ANOVA Results Using Bias as the Criterion

Source	Sum of Squares	*df*	Mean Square	F	*p*	η^2
MODLEN	0.2108	1	0.2108	580.290	< 0.001	0.0958
CONFIG	0.0004	2	0.0002	6.131	0.002	0.0002
DIFF	0.0014	2	0.0007	2.154	0.071	0.0006
COR	0.0005	1	0.0005	1.649	0.139	0.0002
SP	1.3008	2	0.6504	2168.072	< 0.001	0.5913
MODLEN:CONFIG	0.0892	2	0.0446	148.667	< 0.001	0.0405
MODLEN:DIFF	0.1328	2	0.0664	221.318	< 0.001	0.0604
MODLEN:COR	0.0325	1	0.0325	108.072	< 0.001	0.0148
MODLEN:SP	0.1073	2	0.0537	178.835	< 0.001	0.0488
CONFIG:DIFF	0.0005	4	0.0001	0.419	0.757	0.0002
CONFIG:COR	0.0001	2	0.0001	0.167	0.359	0.0000
CONFIG:SP	0.0039	4	0.0010	3.250	0.035	0.0018
DIF:COR	0.0007	2	0.0004	1.167	0.084	0.0003
DIF:SP	0.0251	4	0.0063	20.916	< 0.001	0.0114
COR:SP	0.0093	2	0.0047	15.500	< 0.001	0.0042
Error	0.0072	24	0.0003			

Note: The names of five factors have been simplified: MODLEN = Module Length; CONFIG = Panel Configuration; DIFF = Mean Item Pool Difficulty; COR = Correlation between Item Difficulty and Item Time Intensity; and SP = Speeded Response Pattern.

Table 4.6 Fixed-Effect ANOVA Results Using RMSE as the Criterion

Source	Sum of Squares	*df*	Mean Square	F	*p*	η^2
MODLEN	0.9766	1	0.9766	4883.151	< 0.001	0.1369
CONFIG	0.2193	2	0.1097	548.290	< 0.001	0.0307
DIFF	0.1201	2	0.0600	300.216	< 0.001	0.0168
COR	0.0017	1	0.0008	4.198	0.025	0.0002
SP	2.9342	2	1.4671	7335.621	< 0.001	0.4114
MODLEN:CONFIG	0.3382	2	0.1691	563.617	< 0.001	0.0474
MODLEN:DIFF	0.2163	2	0.1082	360.500	< 0.001	0.0303
MODLEN:COR	0.0805	1	0.0805	268.312	< 0.001	0.0113
MODLEN:SP	1.3955	2	0.6978	2325.839	< 0.001	0.1957
CONFIG:DIFF	0.0337	4	0.0084	28.083	< 0.001	0.0047
CONFIG:COR	0.0081	2	0.0041	13.501	< 0.001	0.0011
CONFIG:SP	0.4412	4	0.1103	367.685	< 0.001	0.0619
DIF:COR	0.0003	2	0.0002	0.521	0.487	0.0000
DIF:SP	0.0117	4	0.0029	9.750	0.001	0.0016
COR:SP	0.0085	2	0.0043	14.161	< 0.001	0.0012
Error	0.0048	24	0.0002			

Note: The names of five factors have been simplified: MODLEN = Module Length; CONFIG = Panel Configuration; DIFF = Mean Item Pool Difficulty; COR = Correlation between Item Difficulty and Item Time Intensity; and SP = Speeded Response Pattern.

Table 4.7 Fixed-Effect ANOVA Results Using MAE as the Criterion

Source	Sum of Squares	*df*	Mean Square	F	*p*	η^2
MODLEN	0.2755	1	0.2755	918.333	< 0.001	0.0880
CONFIG	0.0058	2	0.0029	9.684	< 0.001	0.0019
DIFF	0.0033	2	0.0017	5.517	< 0.001	0.0011
COR	0.0012	1	0.0006	2.036	0.041	0.0004
SP	1.1255	2	0.5628	1875.9	< 0.001	0.3594
MODLEN:CONFIG	0.0734	2	0.0367	122.904	< 0.001	0.0235
MODLEN:DIFF	0.0982	2	0.0491	154.160	< 0.001	0.0314
MODLEN:COR	0.0522	1	0.0522	163.893	< 0.001	0.0167
MODLEN:SP	0.138	2	0.0690	216.641	< 0.001	0.0441
CONFIG:DIFF	0.0008	4	0.0002	0.628	0.371	0.0003
CONFIG:COR	0.0002	2	0.0001	0.314	0.679	0.0001
CONFIG:SP	0.0015	4	0.0004	1.177	0.072	0.0005
DIF:COR	0.0018	2	0.0009	2.826	0.001	0.0006
DIF:SP	0.0582	4	0.0146	45.683	< 0.001	0.0186
COR:SP	0.0071	2	0.0036	11.146		0.0023
Error	0.0072	24	0.0003			

Note: The names of five factors have been simplified.

4.3.2 MST Panel Configuration Characteristics

Each of the panel configuration characteristics was examined to investigate if the levels in them resulted in changes in ability recovery across the theta scale. The panel configuration characteristics included in this section are module length and panel design. Figures 4.5 and 4.6 shows the bias, RMSE, and MAE conditioned on each of the 31 quadrature points across the theta scale.

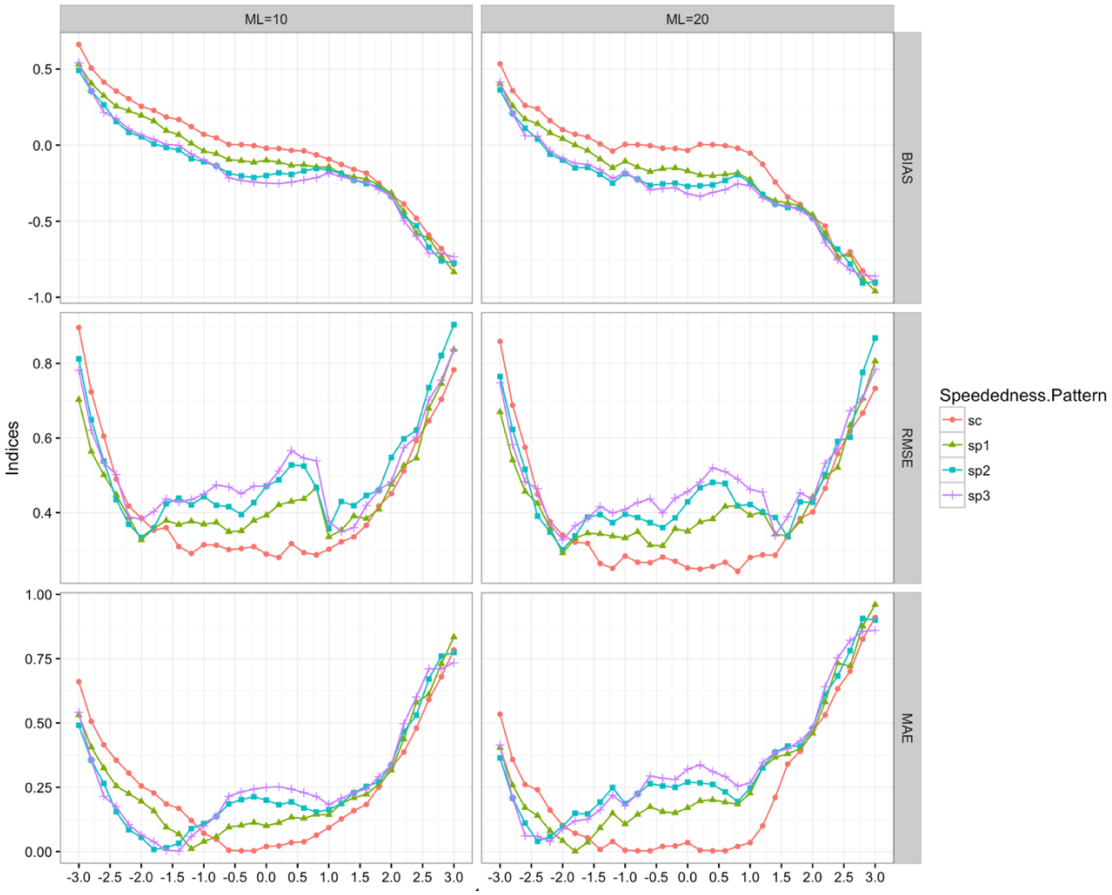


Figure 4.4. Evaluation indices curves across speeded response patterns (bias, RMSE, MAE) by module length.

Module Length

For each of the two module lengths specified in the simulation design, we compared examinees' estimated ability score (theta values) to their true proficiency level under each of the four speeded response patterns. The two levels of module length were short module length (ML = 10) and a medium module length (ML = 20).

As can be seen from the first row of Figure 4.4, all speeded conditions tended to produce overestimated scores for low ability examinees while underestimated high ability examinees' scores. The bias curve of speededness control group is distinctively different from the curves of the other three speeded response patterns at the lower and middle region of the scale. The maximum difference in bias tends to range from 0.15 to 0.27. At the higher region of the theta scale, however, there is little difference in bias across the four bias curves. Within three curves that include speeded responses, the SP2 and the SP3 group produced larger overall biases than did the SP1 group and the difference in bias curve between the SP2 and SP3 groups are negligible. After increasing the module length from 10 to 20, a number of changes need to be noted regarding the four bias curves. First, the estimation accuracy of the proposed speeded control improved at the middle theta scale range as tests become longer. Specifically, the longer test produced satisfactory estimation accuracy from -1 to 0.8. Similar information is also represented in MAE graphs at the third row of Figure 4.4. Secondly, for the three speeded response curves, extending the test length doesn't give rise to the ability estimation accuracy, which is caused by the increased speeded responses due to the fixed speeded response proportion. That is, more speeded responses were produced in longer tests.

Regarding the difference of magnitude of the RMSE, the speededness control group tended to produce the highest RMSE results compared with the other three speeded response groups at the lower range of the theta scale. While this result may be surprising, it's still explainable since the score estimates at the lower theta range are intrinsically overestimated in the traditional MST model, for those low ability examinees in three speeded response groups, however, their scores were lowered by the effect of the speeded response. Therefore, the underestimated scores were corrected in the simulated administration. Although the effect of test speededness was diminished at the lower theta range, its impact on RMSE was more salient at the middle range because the score estimation of intermediate examinees become more accurate, which makes the underestimated scores in speeded response group non-negligible. Within the three speeded response groups, the SP3 condition produced slightly larger biases than did the SP2 condition, which produced larger biases than did the SP1 condition. Moreover, a noticeable decrease of RMSE curves was found for three speeded response groups, indicating a score threshold that examinees with higher scores were not affected by varied test speededness conditions. Overall, the proposed speededness control group outperformed the other three groups with varied levels of speeded responses in terms of RMSE.

Panel Configuration

The other panel characteristic included in this study is the panel configuration design. The three levels of panel configuration design chosen for this study were 1-3 design, 1-2-3 design, and 1-3-3 design. Similar to the module lengths, these MST panel

configurations were compared on the basis of bias, RMSE, and MAE. Figure 4.5 illustrates the differences in these three statistics by panel configuration.

As shown in Figure 4.5, the speededness control group produced better overall bias statistics in three panel configuration designs. A clear gap is shown between the speededness control model and the other three groups. In particular, the SP1 group was found to produce considerably higher bias than the speededness control model, while the bias difference between SP2 and SP3 groups was insignificant. The SP3 group wasn't included in the 1-3 panel design because it requires at least three stages to produce 30% of the responses in both current and the following module. Similar to the module length, configuration designs with modules were associated with better bias estimation for the speededness control group, which means the more items an examinee sees the better the MST model will recover the proficiency score. There was a rather noticeable recovery improvement in terms of bias statistics from the 1-2 design to the 1-2-3 design as an extra stage improved the MST model. The improvement from 1-2-3 to 1-3-3 was not as distinct.

The second row of Figure 4.5 highlights the difference in RMSE by configuration. Similar to the relationship found in module length, there is a decrease in RMSE associated with the increase of the number of modules used in the MST administration. Besides the similar RMSE curve pattern observed in Figure 4.4, the magnitude of negative impact of speeded response pattern on RMSE was dramatic as the number of stages and the number of modules increased in panel configuration, the maximum RMSE in the speeded response group increased from 0.48 to 0.59. The estimation precision for the speededness control group, however, improved as an additional stage added in the

panel configuration, the gap of RMSE curves between speededness control and speeded response group was widened.

4.3.3 Item Pool Characteristics

To determine to what extent item pool characteristics caused changes in the overall ability recovery across four speeded response patterns. Two item pool characteristics, mean item pool difficulty and the parameter correlation between item difficulty and item time intensity were tested in the present study. However, the results of multi factor ANOVA indicated that the mean bias of aggregated ability score estimates were not significantly different as the correlation between difficulty and time intensity changed from 0 to 0.33. Thus, the following section will mainly focus on the relationship between item pool difficulty and the proficiency score recovery in MST.

Mean Item Pool Difficulty

Three item pool difficulty means were chosen to observe the effect of different item pool difficulties on the MST model performance. The three levels of mean difficulty were low difficulty pool (D1 = -1.0), an intermediate difficulty pool (D2 = 0), and a high difficulty pool (D3 = 1.0). Figure 4.6 depicts the evaluation indices by mean difficulty of the item pool.

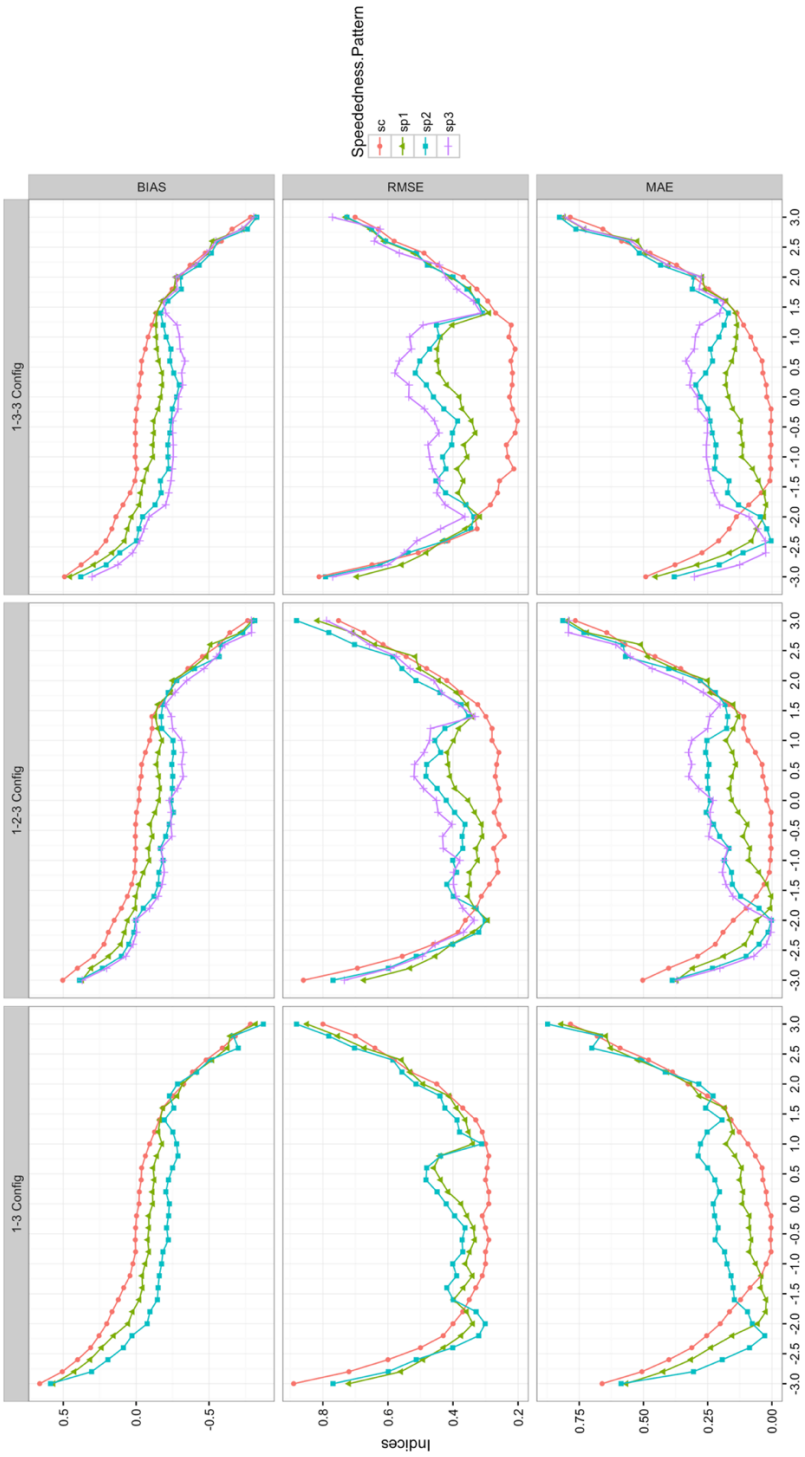


Figure 4.5. Evaluation indices curves across speeded response patterns (bias, RMSE, MAE) by panel configuration.

Since the mean of the true examinee proficiency score distribution was fixed at 0, it is not surprising to see that the MST performed best in terms of score recovery bias when the item pool difficulty was centered at 0. In addition, the optimal theta range for recovering scores also shifted with the center of item pool difficulty. In particular, the proposed speededness control model performs best within the theta range of -1.5 to -0.3 when the center of item pool difficulty is -1. Likewise, the model performs best within the theta range of 0.5 to 1.5 when the pool difficulty is centered at 1. Regarding the bias curves of four speeded response groups, it is apparent that the resemblance of the bias curves of four test speeded response patterns were vastly different across three overall item pool difficulties. In a MST administration with an easy item pool, the majority of examinees were routed to the middle or high level modules, which reduces the amount of speeded response substantially. Therefore, the difference of bias between the four speeded response patterns was negligible. Furthermore, increasing numbers of examinees were routed to the lower level module as the test becomes more challenging. Thereby generating the performance gap of bias index between the speededenss control model and other speeded response pattern groups.

In the same way, the RMSE curves of the four speeded response patterns (SC, SP1, SP2, and SP3) also reflected the relationship between item pool difficulty and the proportion of examinees routed to the lower level module. In addition, the RMSE curves are generally better when item pool difficulty is centered at 0 compared with other RMSE curves in off-centered item pools. Another interesting finding in the RMSE plot is that when the overall test difficulty is moderate (D2 = 0), each of the speeded response curves achieved an equally good RMSE in the lower theta range compared with the speededness

control group. Particularly, as the proportion of speeded response increases, that is, the test overall difficulty increased (D3 = 1.0), the optimal theta estimation range moves further from the speededness control group. In other words, given the simulation results, within the speeded response groups, the proposed program did a better job in estimating low ability examinees when more examinees were routed to lower level modules.

4.4 Item Pool Usage

The item pool usage is reported in this section to depict the relationship between available item pool supplies and the demands of the MST panel configuration as well as the quality of the assembled parallel panels. In the present study, the sizes of assembled item pools were designed in a way that item pools with varied levels of availabilities are employed. Therefore, the impact of item pool size on pool usage can be compared. As a result, some item pools had more items than necessary in certain simulation conditions while other item pools didn't provide the ATA engine with the ideal item supplies. As described in the method chapter, item pool sizes were ideal for panel configurations that use module length of 10. For panel configurations using module length of 20, the degree of item supply insufficiency increases as more stages and modules were added to the configuration. In particular, compared with the ideal item pool sizes, three panel designs (1-3, 1-2-3, 1-3-3) were provided with item pools that have 87%, 60% and 50% of the ideal pool sizes respectively.

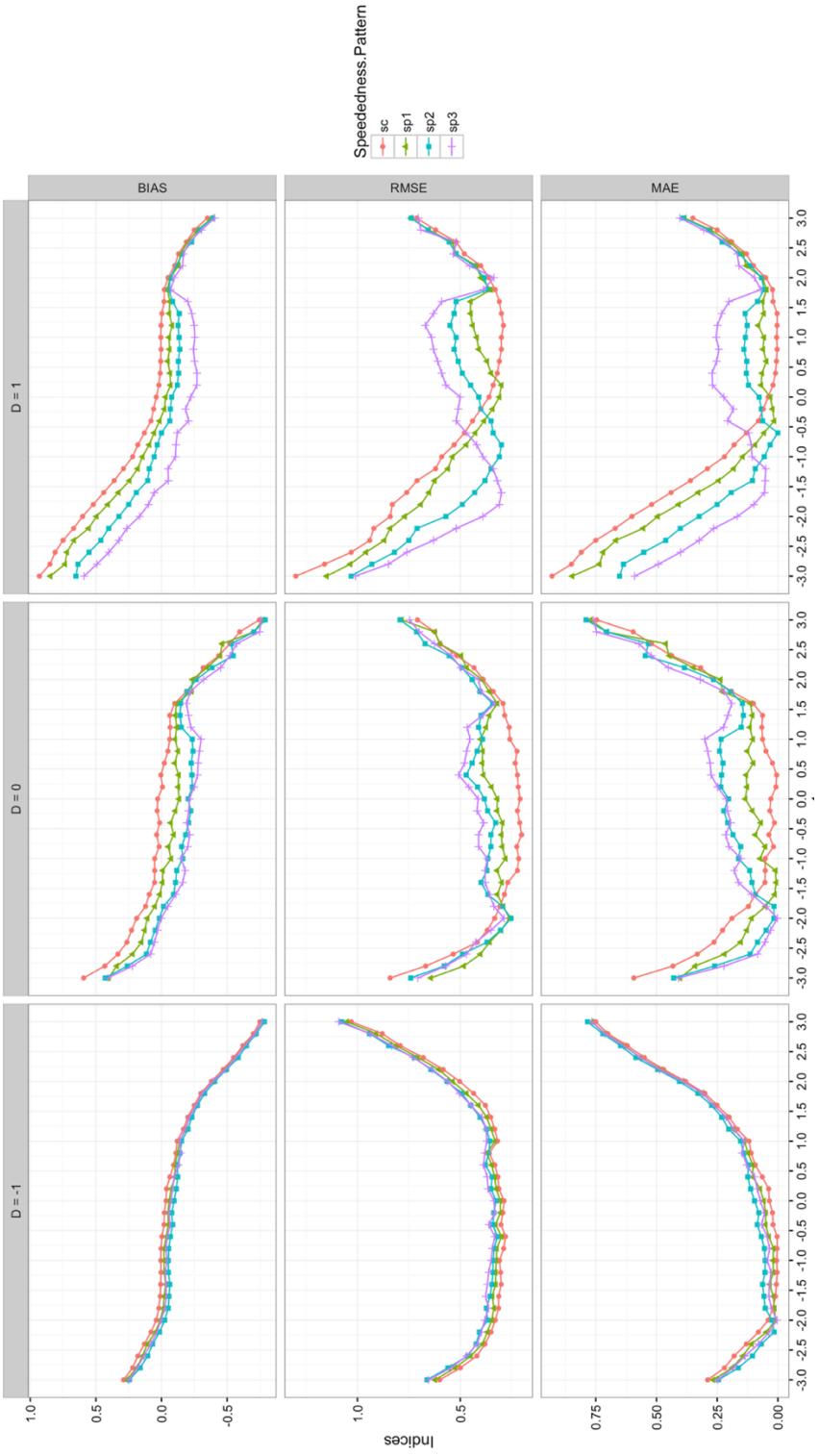


Figure 4.6. Evaluation indices curves across speeded response patterns (bias, RMSE, MAE) by mean item pool difficulty.

Table 4.7 presents the item pool usage evaluation indices for all conditions between panel configuration and item pool characteristics. For each simulation condition, the evaluation indices were reported for both the speededness control group and the averaged speeded response pattern groups. For the overall item pool usage, the traditional MST models that contain speeded response patterns maintained fairly similar item pool usage rates across panel design when the item pool size is ideal (MODLEN = 10). Particularly, 45% to 48% of the items were chosen to assemble the parallel panels. Unsurprisingly, the correlation between item difficulty and item time intensity has no influence on the pool usage rate variation. The speededness control groups, however, used higher proportions of items in the 1-2-3 and 1-3-3 panel design, about 55% and 63% respectively. Moreover, the impact of the correlation between item difficulty and item time intensity was observed. Particularly, when the item difficulty is unrelated to the item time intensity, the proportion of item pool needed to assemble the test is less than the number of items required when item difficulty and item time intensity are correlated.

The second half of Table 4.7 completes the story by describing the item pool usage across panel designs when the item pool supply was not ideal. Specifically, the traditional MST models with speeded response pattern requires a higher proportion of the item pool as the degree of item insufficiency increases from panel design 1-3 to 1-3-3. Nearly 76% of the item pool was used to assemble 10 parallel panels for 1-3-3 when the actual pool size was only 50% of the ideal pool size. A similar increasing trend of item pool usage was also found for the speededness control group where almost 86% of the item pool was used in the lowest item pool supply scenario.

Table 4.8 Comparison of Item Usage Evaluation Between Simulation Conditions

Simulation Conditions				Item Pool Usage				
MODLEN	CONFIG	COR	SP	Item pool size	% of Item pool used	Max item exposure	Mean item exposure	SD of item exposure
10	1-3	0	SP	600	45.44%	0.122	0.084	0.018
		0.33			45.17%	0.119	0.088	0.015
		0	SC		36.54%	0.145	0.113	0.016
		0.33			41.72%	0.133	0.105	0.014
	1-2-3	0	SP	900	46.88%	0.111	0.094	0.012
		0.33			47.04%	0.125	0.089	0.015
		0	SC		52.31%	0.147	0.106	0.021
		0.33			55.37%	0.140	0.094	0.023
	1-3-3	0	SP	1050	48.71%	0.127	0.091	0.018
		0.33			48.39%	0.132	0.087	0.025
		0	SC		56.15%	0.151	0.109	0.021
		0.33			62.79%	0.144	0.098	0.015
20	1-3	0	SP	1050	50.97%	0.133	0.103	0.021
		0.33			51.36%	0.131	0.108	0.019
		0	SC		41.60%	0.163	0.123	0.024
		0.33			48.31%	0.157	0.155	0.025
	1-2-3	0	SP	1050	75.39%	0.139	0.115	0.022
		0.33			76.24%	0.141	0.121	0.020
		0	SC		82.77%	0.165	0.136	0.027
		0.33			85.62%	0.160	0.129	0.026
	1-3-3	0	SP	1050	89.12%	0.158	0.134	0.029
		0.33			90.52%	0.162	0.137	0.030
		0	SC		95.15%	0.184	0.148	0.035
		0.33			95.44%	0.179	0.141	0.031

Note: The names of four simulation conditions have been simplified: MODLEN = Module Length; CONFIG = Panel Configuration; COR = Correlation between Item Difficulty and Item Time Intensity; and SP = Speeded Response Pattern.

Another important evaluation index to examine is the mean item exposure. As can be seen from the Table 4.7, the mean item exposure maintained consistent rates across panel design when the item pool size was ideal. In the simulation conditions where item pool supply is insufficient, like the item pool usage percentage, the mean item exposure rate also increased as the degree of insufficiency become greater. In addition, there are

two patterns in the mean item exposure rate that remain consistent across the simulation conditions: a) the mean item exposure rate of the speededness control model is always higher than the traditional MST with speeded response patterns; b) in the speededness control model, the exposure rate is always higher when the item difficulty is unrelated with the item time intensity.

4.5 Summary

This section of the dissertation presented results from the Automated Test Assembly (ATA) procedure and the MST administration simulation across various conditions. These obtained results directly target the two research questions posed in the previous chapter. Two panel configuration variables were simulated and compared in terms of bias, RMSE, and MAE. (see Figures 4.4 and 4.5). The item pool characteristics question was answered by the plots depicting the effect of item pool characteristic (mean item pool difficulty) on the quality of the proposed speededness control MST model as well as the comparison between the performance of the proposed model and the traditional MST model with varied levels of speeded response patterns.

Overall, the simulation results indicated that the speededness control MST design was superior to the traditional MST with speeded responses in recovering examinees ability score in various panel configurations and item pool characteristic conditions. On the other hand, this study also provided important evidence to advance the understanding of the degree of ability estimation bias in MST administration when speeded responses are present. Additionally, the observed relationship between the module length and the RMSE are consistent with previous simulation studies on MST (e.g., Jodoin et al., 2006).

That is, more items administered generally results in better psychometric accuracy. Likewise, when the panel configuration has more stages, the RMSE dropped due to the increase in test length. Although these results are not surprising, it's useful information for testing agencies to consider if they are planning a transition to MST and need to determine the panel configuration for their MST frameworks.

CHAPTER V

DISCUSSION

This chapter focuses on the implications of the results presented in Chapter IV and situates the discussion in the realm of research introduced in Chapters I and II. The section begins by discussing the characteristics of the speededness control MST model and the effect of speeded response pattern in MST administration. It then focuses on the implications of the simulation results for each of the two research questions. Lastly, it discusses the next steps for future research. It needs to be noted that the obtained simulation results that reflect the role of test speededness in the MST administration is determined by the prespecified simulation operationalization. An operationalization which may not represent the exact real-world scenario when the speededness is presented. The interpretation of the obtained results only applies to the comparison between the varied simulated speeded response patterns and the proposed speededness control model.

5.1 Characteristics of Speededness Control and Speeded Response Panel Configurations

To date, the research focus of MST had revolved around the interactions between the panel configuration demands and the distribution of item pool information and the impact of routing on score precision. This study confirmed some of the findings from previous research, most notably the effect of extending the test length by increasing module length or adding stages. That is, a longer test can lead to improved examinee's ability estimation. In addition, this study took these MST studies a step further and

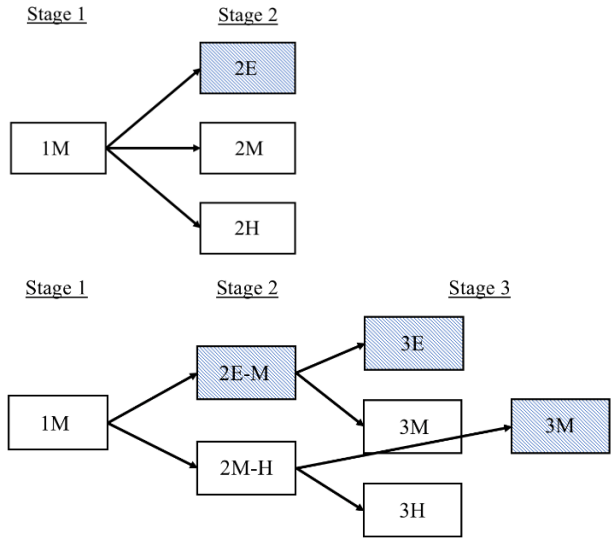
expanded the literature base of the effect of test speededness control in adjusting for the adverse impact of speeded responses in score estimation.

Any potential impact on score estimation due to test speededness can be a real concern, however, testing agencies are not completely free from this matter unless they allow unlimited time. This study identifies a potential source of test speededness in MST administration from the relationship between examinee's ability and the routing feature in the MST design. Two types of test administrations, traditional MST with speeded responses and the speededness control model, were simulated to demonstrate the feasibility of the proposed MST design as well as the necessity of taking account of speeded responses in operational testing administration within the MST framework. Based on the obtained results from the present simulation studies, three aspects of the proposed model are discussed as follows: a) the examinee groups that the proposed MST administration acts on; b) the magnitude of estimation precision improvement; c) the high item pool exposure.

By nature, the adverse effect of test speededness caused by MST routing mainly affects the lower and intermediate levels of examinees since high ability students are not likely to be routed to lower level modules. However, the proportion of examinees being affected by test speededness can vary considerably depending on the choice of panel configuration and simulated speeded response patterns. As can be seen in Figure 4.7, more than 30% of the examinees may produce speeded responses under the effect of test speededness in a 1-3 MST administration, while three out of four paths in a 1-2-3 panel design may contain modules with potential speeded responses. By implementing the

speededness control module design, up to 75% of the potential speeded responses can be adjusted by constructing the parallel modules with controlled test speededness.

Apart from the panel design, the proportion of examinees affected by test speededness is also moderated by the speeded response pattern. That is, certain simulated speeded response patterns may lead to higher proportion of examinees experiencing the test speededness. As introduced in the method chapter, the simulated effect of the test speededness during test-taking experience not only results in speeded responses within a module, it can also lead to speeded responses in the module of the following stage. Regardless of routing direction between stage 2 and stage 3, it is simulated that examinees who are routed to an equal or a higher-level module in stage 3 may still experience the aftermath of routing-caused speededness from stage 2. Although we imposed a rather severe degree of speededness on the present simulation, its impact from the results was substantially large compared to the speeded control condition to indicate that adverse impact of speeded responses on MST score estimation.



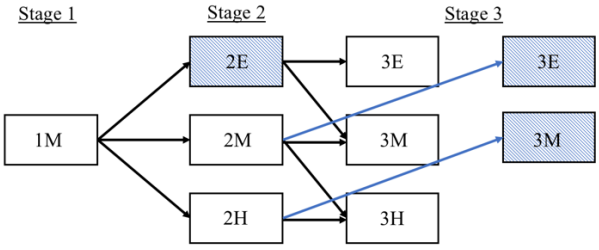


Figure 4. 7 Illustration of Speededness Control Panel Design

Next, the magnitude of estimation precision improvement was illustrated by Figures 4.4 to 4.6, where the graphs clearly showed that the discrepancies between speededness control model (SC) and the traditional MST design with speeded responses were not consistent across the theta scale. In particular, in the lower theta range, the traditional MST designs did a slightly better job in ability score recovery. The speededness control model, on the other hand, noticeably outperformed the traditional MST design in the middle theta range. Lastly, the traditional MST models and the proposed speededness control model performed quite similarly in the higher theta range.

This recovery discrepancy could be attributed to the score underestimation for low and intermediate level examinees in the traditional MST design. Since the scores of examinees with lower thetas tended to be overestimated in MST, the speeded response-caused score underestimation could potentially route more examinees to lower level modules, which exposes them to more low difficulty items. Thus, the MST with speeded responses achieved better estimation precision at the lower end of the theta scale. When it comes to the middle range of the theta scale, more intermediate examinees were erroneously routed to the lower level module which pulls the overall scores down. Furthermore, as depicted in the graph, high ability examinees were untouched by the adverse effect of test speededness since they weren't routed to lower level modules at all

in the administration. Along with the untouched high ability examinee group, there is a decrease of RMSE in the mid and high range of the theta scale. The lowest point of the RMSE curve drop represents the threshold indicating whether or not an examinee was routed to lower level module throughout the course of the administration. The observed drastic change in RMSE within a relatively small theta range runs counter to one of the fundamental requirements in MST administration, that is, score estimation precision should not be related with the routing strategy used in the test. In the speededness control condition, however, the RMSE change was not observed within the same theta range.

Lastly, the high item pool exposure after employing the speededness control MST design is also worth noting. As can be seen in Table 4.7, the MST administration with speededness control tended to use higher proportions of the item pool compared with traditional MST in 1-2-3 and 1-3-3 designs while it used a smaller number of items in the 1-3 design. This difference can be explained by the fact that additional module(s) assembly is required based on the panel design in order to differentiate the routing paths. Thus, 1-2-3 and 1-3-3 designs in speededness control MST models require higher item pool usage rates. Moreover, for the speededness control condition, the consistent lower item pool exposure rates were obtained when the item difficulty is unrelated to the item time intensity. This could be attributed to the simulated artifact that fewer items with low time intensity were available to module assembly when the difficulty is unrelated to the time intensity. Contrastingly, more items are available with both low time intensity and low difficulty for the module assembly when these two parameters are correlated. Overall, although the speededness control condition used more items than traditional MST, the difference was not substantial in item pools with ideal sizes. However, the

absolute pool usage rate was startling for the proposed MST design when the pool size was severely limited, which also highlights its dependence on large size item pools.

5.2 Impact of Panel Configuration

This section addresses the implication of the first research question using the obtained simulation results. Specifically, the research question asked "*To what extent does the proposed MST design recover the simulated proficiency scores under a variety of panel configuration conditions*"? In other words, which panel configuration tends to improve the psychometric estimation accuracy of the proposed speededness control MST model?

The most pronounced difference caused by panel configuration was the evaluative indices (bias and RMSE) tended to improve as the number of stages increased from two to three. As expected, the increase in the MST stage results in more items being added to the test and more items equals better estimation precision. One difference needs to be noted regarding the difference between the proposed speededness control and the traditional MST model is that the estimation accuracy of the proposed model improved as modules were added to the panel configuration whereas no obvious improvement in accuracy was observed in the traditional model with speeded response patterns. A possible explanation for this may be the increase of speeded response since the proportion of speeded responses was fixed regardless of the test length. Thus, longer test lengths also brought more speeded responses which restricts the improvement of accuracy.

Compared with the accuracy improvement (bias, RMSE, and MAE) from the 1-3 to the 1-2-3, the accuracy improvement was much less for the panel design switch from 1-2-3 to 1-3-3. This may indicate that if the pool is sufficiently large and variable for a three-stage design, extending the test length with more modules and making the stage have more difficulty levels doesn't seem to help precision.

The other panel configuration, module length, also made considerable impact on the psychometric quality of the proposed speededness control model. All three evaluation indices improved when the module is 20 items instead of 10 items. The reason for this is exactly the same as adding additional modules within the MST configuration, more items can be translated to better precision. The magnitude of precision improvement from module length extension was about the same as increasing the number of stages.

These two panel configuration variables (number of stages and module length) discussed above seemed to prove a common trend using the simulated data, more items are better. Longer test length are likely to result in increased precision which leads to improved proficiency score recovery. This seems to be true for the proposed MST model. However, the strength of this relationship could be attenuated given the simulated effect of the speeded response patterns. These results imply that testing agencies not only need to ensure that the appropriate number of items are in the module to secure proper estimation precision, but also should develop an updated MST framework that can counter the adverse effect of testing speededness.

5.3 Impact of Item Pool Characteristics

This section discusses the implications of the MST simulation results on the relationship between item pool characteristics and the performance of the proposed MST model. the second research question asked, "*How do various item pool characteristics impact the quality of the proposed MST design*"? This question delves into the relationship between overall item pool characteristics versus panel quality under the demands of the proposed MST configuration.

Between two factors introduced in the simulation study, average item pool difficulty and the correlation between item difficulty and item time intensity, the one that ended up having the significantly larger effect on measurement precision was the average item pool difficulty. Although the correlation between difficulty and time intensity didn't produce significant influence on estimation improvement, its impact on item pool usage was observed. When the item difficulty is related to the item time intensity, that is, easier items tend to require less time to solve. More items were available for ATA engine to choose from in assembling lower level modules, thereby avoiding the situation that a few easy items with low time intensity being overly exposed.

Regarding the average item pool difficulty, this study examined the effect of three average difficulties, D1 = -1, D2 = 0, and D3 = 1. Overall, given the fixed ability proficiency score distribution, N (0, 1), the mean item difficulty parameter of 0 resulted in much better measurement precision than the other two conditions where the mean item difficulty was off-center. Again, this result is not surprising, but the magnitude of improvement was notable. When the item difficulty distribution was largely overlapped

with examinee's ability distribution, it tended to provide more information not only at the center theta range but also more information across the entire theta scale.

As discussed in Chapter IV, the difference of measurement precision between the speededness control group and the traditional MST designs increased as the mean item difficulty shifted from -1 to 1. Moreover, the difference was almost negligible when a relatively easy item pool was used. It suggested that the key component of the superiority of the proposed speededness control model over the traditional MST design is the relative difficulty of item pool to the examinee's ability. This finding directly speaks to the operational tests that are developed for normative reference uses, such as tests designed for awarding additional credentialing to a group of advanced examinees. With higher proportion of examinees likely to find the test challenging, the speeded response factors must be considered and the speededness control design should be employed to secure relative low measurement error for high ability examinees.

In summary, decisions on the characteristics of the item pool and the panel configuration are under total control of the testing agencies developing the test. The obtained simulation results not only showed that the performance of the MST model may vary significantly across different MST features, but also indicates that the impact of test speededness on the examinee's ability estimation is closely related to the combination of decisions made on item pool design and panel configuration.

5.5 Further Research

This section describes further avenues of research to investigate the effect of speededness in MST based testing administration. This section touches on research topics that may further explore the performance of a speededness control MST model.

The most outstanding inadequacy of the present study is the lack of research on the psychological impact of MST routing on examinee's real-time performance to support the practical use of the proposed speededness control model. While the four speeded response patterns were simulated in this study as an attempt to reflect the examinee's test taking behavior in operational tests, several key aspects of the real time test-taking behavior remain uninvestigated, such as the examinee's perception of module difficulty change during the test administration, the pattern of speeded responses when a difficulty change is perceived, and the possible test-taking strategies examinees may take to deal with the negative feedback. If the aforementioned aspects can be further explored, the practical utility of the proposed MST design may become clearer.

One limitation of the study was that the same speeded response pattern was used throughout the entire MST administration for a given condition. Further investigation can vary the degree of test speededness effect by specifying different levels of speeded response pattern by stage and the starting module level. For instance, an examinee routed from a moderate stage 1 module to an easy stage 2 module may experience a different level of speededness effect from another examinee who takes the route from a difficult stage 2 module to an intermediate stage 3 module. Again, future research on psychological impact of routing on test taking experience is warranted to accurately depict the varied speeded response patterns in the MST administration.

Another limitation in the proposed MST model was its insufficiency of flexibility to combine multiple speeded response patterns into one MST administration. All simulated examinees in a current given condition can only possess one speeded response pattern, while the reality is examinees with varied psychological adaptabilities will take the same test together. One possible solution that future research can investigate is to create an on-the-fly MST that can identify the examinee's speeded response pattern, or his or her psychological adaptability toward module difficulty change, after the first routing and adaptively assemble the subsequent module that match with both the estimated ability and the speeded response pattern.

Lastly, more exotic panel configurations can be added to the simulation condition. The present study sampled three most common conditions out of a very large population of potential MST configurations, while the potential relationship between the proposed speededness control model and the panel configuration complexity had yet to be investigated. For instance, a study could sample configurations such as 1-5-5-5-5 or 1-3-5-7 or other configurations with much larger spread in both stages and difficulty levels.

These are just a few of the many research endeavors that can be done to examine the field of test speededness within the MST framework. These studies will collectively help to inform testing agencies in test design and further the research base of test speededness in the operational use of the MST framework.

REFEENCES

Armstrong, R., Jones, D., Koppel, N. & Pashley, P. (2000). *Computerized adaptive testing with multiple forms structures.* Paper presented at the meeting of the National Council on Measurement in Education, New Orleans, LA.

Ariel. A., van der Linden. W. J., & Veldkamp. B. P. (2006). A strategy for optimizing item-pool management. *Journal of Educational Measurement, 43,* 85-96

Bejar, I. I. (1985). *Test speededness under number-right scoring: An analysis of the Test of English as a Foreign Language* (Report No. ETS-RR-85-11). Princeton, NJ: Educational Testing Service. (ERIC Document Reproduction Service No. ED 263128)

Bergstrom, B., Lunz, M., & Gershon, R.C. (1994) An empirical study of computerized adaptive test administration conditions. *Journal of Educational Measurement, 31,* 251-263.

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

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2002). Item parameter estimation under conditions of test speededness: Application of a mixture Rasch model with ordinal constraints. *Journal of Educational Measurement, 39*, 331-348.

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*, 1061–1071.

Breithaupt, K., Ariel, A., & Hare, D. R. (2010) Assembling an inventory of multistage adaptive testing systems. *Elements of Adaptive Testing,* 247-266.

Bridgeman. B., & Cline. F (2004) Effect of differentially time-consuming tests on computer-adaptive test scores. *Journal of Educational Measurement, 41,* 137-148.

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

Chang, H-.H. (2015) Psychometric behind computerized adaptive testing. *Psychometrika. 80.* 1-20.

Chen, J. (2004) Effects of test anxiety, time pressure, ability and gender on response aberrance. Unpublished doctoral dissertation, The Ohio State University.

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.

Choe, E. M., & Kern, J. L. (2014). *Controlling item exposure for response time-informed item selection in CAT.* Paper presented at the annual meeting of the American Educational Research Association, Philadelphia, PA

Cline. F., & Powers., D (2014) Test-taker perceptions of the role of the GRE general test in graduate admissions: preliminary findings. *The Research Foundation for the GRE revised General Test: A Compendium of Studies.*

Cohen, A. S., Wollack, J. A., Bolt, D. M., & Mroch, A. A. (2002, April). *A mixture Rasch model analysis of test speededness*. Paper presented at the Annual Meeting of American Educational Research Association, New Orleans, LA.

Colwell. N (2013) Test anxiety, computer-adaptive testing and the common core. *Journal of Education and Training Studies, 1,* 50-60.

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

Elliott, S. N., & McKevitt, B. (2000, April). *Experimental analysis of the effects of testing accommodations on the scores of students with disabilities: Design issues and initial results.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans

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

Fan, Z., Wang, C., Chang, H.-H., & Douglas, J. A. (2012). Utilizing response time distributions for item selection in CAT. *Journal of Educational and Behavioral Statistics*, *37,* 655-670.

Ferrando, P. J., & Lorenzo-Seva, U. (2007). An item-response model incorporating response time data in binary personality items. *Applied Psychological Measurement, 31*, 525–543.

Frey, A., Hartig, J., & Moosbrugger, H. (2009). Effects of adaptive testing on test taking motivation. *Diagnostica, 55*, 20-28.

Gaviria, J.-L. (2005). Increase in precision when estimating parameters in computer assisted testing using response times. *Quality & Quantity, 39,* 45–69.

Geen, R. G. (1987). Test anxiety and behavioral avoidance. *Journal of Research in Personality, 21,* 481-488.

Gulliksen, H. (1950). *Theory of mental tests*. Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Hausler. J., & Sommer. M. (2008) The effect of success probability on test economy and self-confidence in computerized adaptive tests. *Psychology Science, 50,* 75-87.

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

Hill, K. T., & Eaton, W. O. (1977). The interaction of test anxiety and success-failure experiences in determining children's arithmetic performance. *Developmental Psychology, 13,* 205-211.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Hong, E. (1998). Differential stability of individual differences in state and trait test anxiety. *Learning and Individual Differences, 10,* 51-69.

Hornke, L. F. (2000). Item response times in computerized adaptive testing. *Psicologica*, *21,* 175–189.

Jenkins, S. M., & Holmes, S. D. (1999). *Computer usage and access patterns of actual and potential LSAT takers*. (Computerized Testing Report 97-10). Newtown, PA: Law School Admission Council.

Jodoin, M. G. (2006) Comparison of the psychometric properties of several computer-based test designs for credentialing exams with multiple purposes. *Applied Measurement in Education, 19,* 203-220.

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

Kimura T, Nagaoka K. *Psychological aspects of CAT: how test-takers feel about CAT.* Proceedings of the International Association for Computer Adaptive Testing Conference; 2011 Oct 3-5; Pacific Grove, USA. Woodbury, MN. 2011.

Klein-Entink, R. H., Fox, J.-P., & van der Linden, W. J. (2009). A multivariate multilevel approach to the modeling of accuracy and speed of test takers. *Psychometrika, 74,* 21-48.

Kong, X. J., Wise, S. L., & Bhola, D. S. (2007). Setting the response time threshold parameter to differentiate solution behavior from rapid-guessing behavior. *Educational and Psychological Measurement*, *67*, 606-619.

Lee, Y., Jia, Y. (2014) Using response time t investigate students' test-taking behaviors in a NAEP computer-based study. *Large-scale Assessment in Education.* 2:8.

Liebert. R. M., & Morris. L. W. (1967) Cognitive and emotional components of test anxiety: a distinction and some initial data. *Psychological Report, 20,* 975-978.

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

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

Lord, F. M. (1983). Unbiased estimators of ability parameters, of their variance, and of their parallel-forms reliability. *Psychometrika*, *48*, 233-245.

Lu, Y., & Sireci, S. (2007) Validity issues in test speededness. *Educational Measurement: Issues and Practices. 26,* 29-37

Luecht, R. M., Nungester, R. J. & Hadadi, A. (1996, April). *Heuristic-based CAT: Balancing item information, content and exposure.* Paper presented at the meeting of the National Council on Measurement in Education, New York.

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

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

Luce, R. D. (1986). *Response times: Their roles in inferring elementary mental organization.* Oxford, UK: Oxford University Press.

Marianti, S., Fox, J., Avetisyan, M., Veldkamp, B., & Tijmstra, J. (2014) Testing for Aberrant Behavior in Response Time Modeling. *Journal of Educational and Behavioral Statistics. 39*, 426-451

Marquart, A. M. (2000, June). *The use of extended time as an accommodation on a standardized mathematics test: An investigation of effects on scores and perceived consequences for students of various skill levels.* Paper presented at the annual meeting of the Council of Chief State School Officers, Snowbird, UT. Retrieved December 12, 2001,

McGlohen, M., & Chang H-.H. (2008) Combining computer adaptive testing technology with cognitively diagnostic assessment. *Behavior Research Method. 40*, 808-821.

Melican, G. J., Breithaupt, K., & Zhang, Y. (2010). Designing and implementing a multisatage adaptive test: The uniform CPA exam. In W. J. van der Linden & C. a. W.

Glas (Eds.), *Elements of adaptive testing* (pp. 167–190). New York, NY: Springer.

Meyer, J. P. (2008). *A mixture Rasch model with item response-time components.* Paper presented at the annual meeting of the National Council on Measurement in Education, New York, NY.

Morse. D., & More. L (2002) Are undergraduate examinees' perceptions of item difficulty related to item characteristics? *Perceptual and Motor Skills, 95*, 1281-1286.

Morris, L. W., & Liebert, R. M. (1969). The effects of anxiety on timed and untimed intelligence tests: another look. *Journal of Consulting and Clinical Psychology, 33,* 240-244.

Orfus. S. (2008) The effect test anxiety and time pressure on performance. *The Huron University College Journal of Learning and Motivation, 46,* 118-133.

Oshima, T. C. (1994). The effect of speededness on parameter estimation in item response theory. *Journal of Educational Measurement, 31*, 200–219.

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

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

Phillips, S. E. (1994). High-stakes testing accommodations: Validity versus disabled rights. *Applied Measurement in Education, 7(2)*, 93–120.

Pitkin, A. K., & Vispoel, W. P. (2001). Differences between self-adapted and computerized adaptive tests: A meta-analysis. *Journal of Educational Measurement, 38(3),* 235-247.

Powers, E. D. (2001) Test anxiety and test performance: comparing paper-based and computer-apaptive versionsof the graduate record examinations Gernal Test. *Journal of Educational Computing Research, 24,* 249-273.

Powers E. D. (2008) Incidence, correlates, and possible causes of test anxiety in graduate admission testing, *Advances in Personality Assessment, 7,* 49-75.

Ragosta, M., & Wendler, C. (1992). *Eligibility issues and comparable time limits for disabled and nondisabled SAT examinees* (College Board Research Report No. 92-5). New York: The College Board.

R Core Team (2014) *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

Ranger, J., Kuhn, J. T., & Gaviria, J. L. (2015). A Race Model for Responses and Response Times in Tests. *Psychometrika, 80,* 791–810.

Rindler, S. E. (1979). Pitfalls in assessing test speededness. *Journal of Educational Measurement, 16*, 261-270.

Roskam, E. E. (1997). Models for speed and time-limit tests. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory*. Springer-Verlag.

Rouder, J. N., Province, J. M., Morey, R. D., Gomez, P., & Heathcote, A. (2015). The lognormal race: A cognitive process model of choice and latency with desirable psychometric properties. *Psychometrika, 80,* 491–513.

Sarason. G. (1984) Stress, anxiety, and cognitive interference: Reactions to tests. *Journal of Personality and Social Psychology, 46(4)*, 929-938.

Schnipke, D. L. (1995, April). *Assessing speededness in computer-based tests using item response times*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.

Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement*, *34*, 213-232.

Schnipke, D. L., & Scrams, D. J. (2002). Exploring issues of examinee behavior: Insights gained from response-time analyses. In C. N. Mills, M. Potenza, J. J. Fremer & W. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 237–266). Hillsdale, NJ: Lawrence Erlbaum Associates.

Scheiblechner, H. (1979). Specific objective stochastic latency mechanisms. *Journal of Mathematical Psychology*, *19*, 18-38.

Stretch, L. S., & Osborne, J. W. (2005). Extended time test accommodation: Directions for
future research and practice. *Practical Assessment, Research & Evaluation, 10*, 1–8.

Tatsuoka, K. K., & Tatsuoka, M. M. (1980). A model for incoporating response-time data in scoring achievement tests. In D. J. Weiss (Ed.), *Proceedings of the 1979 computerized adaptive testing conference.*

Tippets, E., & Benson, J. (1989) The effect of item arrangement on test anxiety, *Applied Measurement in Education, 2,* 289-296.

Thissen, D. (1983). Timed testing: An approach using item response theory. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing*. New York, NY: Academic Press.

Van Breukelen, G. J. P. (2005). Psychometric modeling of response speed and accuracy with mixed and conditional regression. *Psychometrika*, *70*, 359-376.

van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory*. New York, NY: Springer.

van der Linden, W. J. (1998). Bayesian item selection criteria for adaptive testing. *Psychometrika*, 63, 201-216.

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

van der Linden, W. J., & van Krimpen-Stoop, E. M. L. A. (2003). Using response times to detect aberrant responses in computerized adaptive testing. *Psychometrika, 68*, 251-265.

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

van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics*, *31,* 181-204.

van der Linden, W. J., Breithaupt, K., Chuah, S., & Zhang, Y. (2007a) Detecting differential speededness in multistage testing. *Journal of Educational Measurement, 44,* 117-130.

van der Linden, W. J. (2007b). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, 72 , 387-308.

van der Linden, W. J., & Guo, F. (2008a). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika*, *73*, 365-384.

van der Linden, W. J. (2008b). Using response times for item selection in adaptive testing. *Journal of Educational and Behavioral Statistics, 33*, 5-20.

van der Linden, W. J. (2009a). Conceptual issues in response-time modeling. *Journal of Educational Measurement*, *46*, 247-272.

van der Linden, W. J. (2009b). Predictive control of speededness in adaptive testing. *Applied Psychological Measurement, 33*, 25-41.

van der Linden, W. J. (2011) Setting time limits on tests. *Applied Psychological Measurement, 35*, 183-199.

van der Maas, H. L. J., Molenaar, D., Maris, G., Kievit, R. A., & Borsboom, D. (2011). Cognitive psychology meets psychometric theory: On the relation between process

models for decision making and latent variable models for individual differences. *Psychological Review, 118,* 339–356.

Veldkamp, B. P., & van der Linden, W. J. (2010) Designing item pools for adaptive testing. *Elements of adaptive testing,* 231-245.

Veldkamp, B. P. (2016). On the issue of item selection in computerized adaptive testing with response times. *Journal of Educational Measurement, 53*, 212-228.

Verhelst, N. D., Verstralen, H. H. F. M., & Jansen, M. G. (1997). A logistic model for time-limit tests. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory.* New York, NY: Springer-Verlag.

Vispoel, W. P. (1998). Reviewing and changing answers on computer adaptive and self-adaptive vocabulary tests. *Journal of Educational Measurement, 35*, 328–345.

Vispoel, W. P., Clough. S., & Bleiler. T. (2002) Can Examinees Use Judgments of Item Difficulty to Improve Proficiency Estimates on Computerized Adaptive Vocabulary Tests? *Journal of Educational Measurement, 39*, 311-330.

Wainer, H. & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24*, 185–201.

Wang, T., & Hanson, B. A. (2005). Development and calibration of an item response model that incorporates response time. *Applied Psychological Measurement*, *29,* 323-339.

Wang, T. (2006). *A model for the joint distribution of item response and response time using one-parameter Weibull distribution* (CASMA Research Report 20). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment.

Wang, X., Fluegge, L., & Luecht, R. M. (2012) *A large-scale comparative study of the accuracy and efficiency of ca-MST panel design configurations.* Presented at the National Council on Measurement in Eudcation, Vancouver, BC, Canada.

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

Willse, J., Ackerman, T., & Luecht R. M. (2012). *An overview of ca-MST: From panel configurations to test assembly*. Paper Presented at National Council of Measurement in Education in Vancouver, BC.

Wise, L. L. (1986, April). *Latent trait models for partially speeded tests.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.

Wise, S. L., Finney, S., Enders, C (1999) Examinee Judgments of Changes in Item Difficulty: Implications for Item Review in Computerized Adaptive Testing. *Applied Measurement in Education*, *12,* 185-198.

Wise, S. L., & DeMars C. E. (2006). An application of item response time: The effort moderated IRT model. *Journal of Educational Measurement, 43*, 19–38.

Wise, S. L., & Ma, L. (2012) *Setting response time thresholds for a CAT item pool: the normative threshold method*. Paper presented at the 2012 annual meeting of the National Council on Measurement in Education. Vancouver, Canada.

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

Yamamoto, K., & Everson, H. (1997). Modeling the effects of test length and test time on parameter estimation using the hybrid model. In J. Rost (Ed.), *Applications of latent trait and latent class models in the social sciences.* 89-98. Münster, Germany: Waxmann.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*, 187–214.

Zeidner, M. (1998*). Test anxiety: The state of the art.* New York: Plenum Press.

Zenisky, A. L., Hambleton, R. K., & Sireci, S. G. (2002). Identification and evaluation of local item dependencies in the Medical College Admissions Test. *Journal of Educational Measurement, 39*, 291–309.

Zenisky, A., Hambleton, R., Luecht, R. M. (2010) Multistage Testing: Issues, Designs, and Research. *Elements of Adaptive Testing. Springer.*

APPENDIX

Table 4. 1 RMSE of Generated MIFs Against Target in Two-Stage Panel Configuration (1-3).

Condition	Stage1 - M	Stage2 - L	Stage 2 - M	Stage 2 - H
C1L1D1R1	0.45	0.41	0.44	0.44
C1L1D1R2	0.44	0.44	0.41	0.42
C1L1D2R1	0.39	0.40	0.37	0.45
C1L1D2R2	0.40	0.39	0.42	0.39
C1L1D3R1	0.43	0.45	0.38	0.41
C1L1D3R2	0.45	0.43	0.41	0.40
C1L2D1R1	0.95	0.97	0.86	0.96
C1L2D1R2	0.93	0.98	0.83	0.99
C1L2D2R1	1.07	1.03	1.11	1.05
C1L2D2R2	1.02	1.09	1.03	0.93
C1L2D3R1	0.98	1.02	1.07	1.10
C1L2D3R2	1.05	0.95	1.08	1.05

Note. Study variable names are abbreviated in the condition column. C, L, D and R denote configuration, module length, average item difficulty and the correlation between item difficulty and time intensity respectively.

Table 4. 2 RMSE of Generated MIFs Against Target in Three-Stage Panel Configuration (1-2-3).

	Stage1 - M	Stage2 - L	Stage 2 - H	Stage 3 - L	Stage 3 - M	Stage 3 - H
C2L1D1R1	0.42	0.45	0.46	0.43	0.44	0.40
C2L1D1R2	0.45	0.43	0.47	0.46	0.47	0.47
C2L1D2R1	0.44	0.41	0.45	0.48	0.47	0.44
C2L1D2R2	0.40	0.42	0.45	0.49	0.43	0.47
C2L1D3R1	0.41	0.41	0.42	0.49	0.51	0.49
C2L1D3R2	0.44	0.45	0.49	0.49	0.49	0.43
C2L2D1R1	1.19	1.2	0.99	1.07	1.04	1.17
C2L2D1R2	0.99	1.12	1.09	1.14	1.03	1.13
C2L2D2R1	1.08	1.18	1.12	1.14	1.13	1.06
C2L2D2R2	1.18	1.01	1.03	1.09	1.08	1.05
C2L2D3R1	0.98	1.07	0.98	1.14	1.03	1.18
C2L2D3R2	1.18	1.17	0.99	1.14	0.99	1.12

Note. Study variable names are abbreviated in the condition column.

Table 4. 3 RMSE of Generated MIFs Against Target in Three-Stage Panel Configuration (1-3-3).

	Stage1-M	Stage2-L	Stage2-M	Stage2-H	Stage3-L	Stage3-M	Stage3-H
C2L1D1R1	0.43	0.48	0.52	0.49	0.43	0.44	0.50
C2L1D1R2	0.5	0.42	0.47	0.47	0.49	0.49	0.53
C2L1D2R1	0.43	0.42	0.47	0.47	0.42	0.51	0.42
C2L1D2R2	0.41	0.53	0.51	0.51	0.41	0.44	0.46
C2L1D3R1	0.44	0.51	0.52	0.42	0.41	0.50	0.49
C2L1D3R2	0.41	0.42	0.48	0.52	0.48	0.44	0.44
C2L2D1R1	1.18	1.06	1.09	1.11	1.12	1.11	1.04
C2L2D1R2	1.05	1.14	1.03	1.1	1.10	1.12	1.09
C2L2D2R1	1.18	1.09	1.01	1.03	1.11	1.13	1.07
C2L2D2R2	1.01	1.03	1.13	1.19	1.16	1.07	1.18
C2L2D3R1	1.07	1.05	1.07	1.14	1.04	1.15	1.12
C2L2D3R2	1.19	1.01	1.13	1.07	1.09	1.08	1.04

Note. Study variable names are abbreviated in the condition column.

Table 4. 4 Proportion of Routing Path Taken by Simulated Examinees in an Easy Item Pool Under a Three-Stage Panel Design (1-2-3)

Rep	M-L-L				M-L-M				M-H-M				M-H-H			
	SC	SP1	SP2	SP3	SC	SP1	SP2	SP3	SC	SP1	SP2	SP3	SC	SP1	SP2	SP3
1	0.14	0.18	0.18	0.1	0.18	0.2	0.3	0.24	0.35	0.22	0.16	0.25	0.36	0.36	0.35	0.38
2	0.15	0.12	0.19	0.11	0.19	0.2	0.33	0.27	0.34	0.22	0.22	0.22	0.32	0.32	0.36	0.38
3	0.16	0.13	0.2	0.13	0.21	0.21	0.26	0.24	0.35	0.24	0.15	0.25	0.32	0.37	0.38	0.4
4	0.11	0.12	0.19	0.12	0.23	0.28	0.28	0.26	0.25	0.24	0.19	0.25	0.36	0.33	0.38	0.39
5	0.09	0.16	0.15	0.18	0.24	0.28	0.26	0.21	0.28	0.2	0.2	0.26	0.4	0.4	0.35	0.32
6	0.07	0.13	0.21	0.18	0.25	0.26	0.28	0.24	0.33	0.27	0.15	0.29	0.32	0.34	0.35	0.4
7	0.16	0.14	0.19	0.1	0.24	0.21	0.31	0.24	0.26	0.3	0.2	0.24	0.3	0.38	0.34	0.32
8	0.09	0.15	0.15	0.2	0.18	0.27	0.32	0.23	0.35	0.24	0.21	0.23	0.38	0.33	0.4	0.4
9	0.01	0.14	0.21	0.16	0.24	0.2	0.29	0.26	0.33	0.2	0.19	0.29	0.35	0.37	0.37	0.37
10	0.07	0.13	0.17	0.14	0.18	0.26	0.31	0.27	0.34	0.22	0.24	0.3	0.32	0.36	0.39	0.34

Table 4. 5 Proportion of Routing Path Taken by Simulated Examinees in an Moderate Item Pool Under a Three-Stage Panel Design (1-2-3)

Rep	M-L-L				M-L-M				M-H-M				M-H-H			
	SC	SP1	SP2	SP3	SC	SP1	SP2	SP3	SC	SP1	SP2	SP3	SC	SP1	SP2	SP3
1	0.18	0.3	0.3	0.27	0.26	0.32	0.33	0.26	0.19	0.19	0.12	0.12	0.18	0.14	0.07	0.17
2	0.21	0.31	0.35	0.33	0.27	0.3	0.36	0.32	0.24	0.14	0.15	0.2	0.18	0.13	0.05	0.12
3	0.22	0.32	0.32	0.27	0.2	0.3	0.39	0.3	0.19	0.13	0.11	0.2	0.21	0.2	0.14	0.19
4	0.24	0.3	0.36	0.3	0.25	0.28	0.34	0.3	0.23	0.16	0.13	0.18	0.23	0.19	0.12	0.17
5	0.22	0.26	0.31	0.29	0.29	0.29	0.34	0.32	0.2	0.2	0.14	0.14	0.24	0.1	0.09	0.2
6	0.2	0.3	0.3	0.34	0.3	0.31	0.38	0.25	0.2	0.12	0.11	0.15	0.19	0.11	0.12	0.17
7	0.23	0.27	0.3	0.28	0.23	0.3	0.4	0.3	0.18	0.2	0.13	0.16	0.22	0.13	0.05	0.18
8	0.25	0.34	0.3	0.3	0.22	0.27	0.3	0.29	0.21	0.18	0.11	0.16	0.23	0.18	0.05	0.14
9	0.23	0.29	0.38	0.26	0.25	0.33	0.35	0.26	0.19	0.14	0.14	0.13	0.19	0.14	0.08	0.19
10	0.22	0.26	0.31	0.34	0.21	0.26	0.34	0.28	0.19	0.2	0.15	0.2	0.25	0.13	0.05	0.18

Table 4. 6 Proportion of Routing Path Taken by Simulated Examinees in an Difficult Item Pool Under a Three-Stage Panel Design (1-2-3)

Rep	M-L-L				M-L-M				M-H-M				M-H-H			
	SC	SP1	SP2	SP3	SC	SP1	SP2	SP3	SC	SP1	SP2	SP3	SC	SP1	SP2	SP3
1	0.46	0.52	0.51	0.45	0.27	0.37	0.32	0.3	0.15	0.18	0.11	0.13	0.17	0.1	0.06	0.04
2	0.45	0.45	0.53	0.55	0.31	0.31	0.33	0.28	0.24	0.23	0.12	0.16	0.11	0.09	0.01	0.06
3	0.46	0.52	0.59	0.45	0.3	0.32	0.34	0.31	0.21	0.17	0.12	0.13	0.15	0.08	0.01	0.06
4	0.44	0.51	0.52	0.53	0.25	0.29	0.4	0.29	0.18	0.16	0.1	0.13	0.13	0.07	0.06	0.01
5	0.46	0.52	0.54	0.55	0.31	0.29	0.36	0.32	0.17	0.22	0.13	0.21	0.1	0.07	0.06	0.02
6	0.4	0.54	0.52	0.55	0.28	0.37	0.4	0.37	0.23	0.15	0.14	0.15	0.13	0.07	0.02	0.06
7	0.43	0.5	0.54	0.49	0.25	0.31	0.36	0.34	0.21	0.13	0.1	0.18	0.11	0.06	0.06	0.05
8	0.36	0.52	0.52	0.46	0.3	0.37	0.34	0.35	0.22	0.23	0.1	0.21	0.13	0.09	0.03	0.04
9	0.37	0.53	0.52	0.54	0.27	0.35	0.38	0.3	0.16	0.19	0.11	0.18	0.16	0.04	0.04	0.03
10	0.43	0.55	0.52	0.54	0.25	0.29	0.33	0.31	0.23	0.21	0.18	0.22	0.11	0.1	0.06	0.04

Table 4. 7 Evaluation Indices of Ability Score Recovery Conditioned on Module Length (ML = 10)

Theta	Bias				RMSE				MAE			
	SC	SP1	SP2	SP3	SC	SP1	SP2	SP3	SC	SP1	SP2	SP3
-3	0.661	0.531	0.491	0.541	0.896	0.703	0.812	0.782	0.661	0.531	0.491	0.541
-2.8	0.506	0.406	0.356	0.356	0.724	0.564	0.649	0.621	0.506	0.406	0.356	0.356
-2.6	0.415	0.325	0.265	0.215	0.605	0.501	0.538	0.534	0.415	0.325	0.265	0.215
-2.4	0.355	0.255	0.155	0.175	0.491	0.448	0.435	0.502	0.355	0.255	0.155	0.175
-2.2	0.305	0.225	0.085	0.105	0.418	0.387	0.369	0.388	0.305	0.225	0.085	0.105
-2	0.255	0.195	0.055	0.065	0.385	0.327	0.334	0.383	0.255	0.195	0.055	0.065
-1.8	0.228	0.158	0.008	0.038	0.353	0.358	0.359	0.403	0.228	0.158	0.008	0.038
-1.6	0.185	0.095	-0.015	0.005	0.36	0.379	0.424	0.437	0.185	0.095	0.015	0.005
-1.4	0.168	0.068	-0.032	-0.002	0.309	0.368	0.439	0.428	0.168	0.068	0.032	0.002
-1.2	0.121	0.011	-0.089	-0.059	0.291	0.377	0.421	0.435	0.121	0.011	0.089	0.059
-1	0.071	-0.039	-0.109	-0.099	0.314	0.369	0.443	0.454	0.071	0.039	0.109	0.099
-0.8	0.048	-0.057	-0.137	-0.137	0.313	0.374	0.42	0.474	0.048	0.057	0.137	0.137
-0.6	0.005	-0.095	-0.185	-0.215	0.301	0.349	0.416	0.469	0.005	0.095	0.185	0.215
-0.4	0.003	-0.102	-0.202	-0.232	0.304	0.351	0.395	0.451	0.003	0.102	0.202	0.232
-0.2	-0.003	-0.113	-0.213	-0.243	0.309	0.379	0.427	0.471	0.003	0.113	0.213	0.243
0	-0.021	-0.112	-0.223	-0.25	0.289	0.393	0.471	0.471	0.022	0.123	0.202	0.253
0.2	-0.022	-0.112	-0.182	-0.252	0.281	0.421	0.488	0.512	0.022	0.112	0.182	0.252
0.4	-0.035	-0.133	-0.193	-0.243	0.317	0.43	0.528	0.566	0.035	0.133	0.193	0.243
0.6	-0.038	-0.129	-0.169	-0.229	0.293	0.437	0.525	0.546	0.038	0.129	0.169	0.229
0.8	-0.064	-0.144	-0.154	-0.214	0.287	0.467	0.467	0.539	0.064	0.144	0.154	0.214
1	-0.093	-0.143	-0.163	-0.183	0.302	0.335	0.357	0.377	0.093	0.143	0.163	0.183
1.2	-0.127	-0.187	-0.187	-0.207	0.322	0.353	0.43	0.349	0.127	0.187	0.187	0.207
1.4	-0.159	-0.209	-0.229	-0.229	0.335	0.391	0.419	0.36	0.159	0.209	0.229	0.229
1.6	-0.183	-0.223	-0.253	-0.243	0.366	0.384	0.446	0.419	0.183	0.223	0.253	0.243
1.8	-0.251	-0.261	-0.271	-0.291	0.418	0.409	0.461	0.461	0.251	0.261	0.271	0.291
2	-0.326	-0.316	-0.336	-0.336	0.451	0.476	0.548	0.483	0.326	0.316	0.336	0.336
2.2	-0.387	-0.437	-0.467	-0.497	0.512	0.526	0.598	0.574	0.387	0.437	0.467	0.497
2.4	-0.483	-0.581	-0.53	-0.623	0.593	0.546	0.621	0.603	0.482	0.581	0.533	0.602
2.6	-0.591	-0.611	-0.671	-0.711	0.646	0.679	0.735	0.702	0.591	0.611	0.671	0.711
2.8	-0.681	-0.732	-0.761	-0.711	0.704	0.746	0.821	0.755	0.682	0.732	0.761	0.711
3	-0.784	-0.834	-0.774	-0.734	0.783	0.837	0.904	0.837	0.784	0.834	0.774	0.734

Table 4. 8 Evaluation Indices of Ability Score Recovery Conditioned on Module Length (ML = 20)

Theta	Bias				RMSE				MAE			
	SC	SP1	SP2	SP3	SC	SP1	SP2	SP3	SC	SP1	SP2	SP3
-3	0.534	0.404	0.364	0.414	0.859	0.67	0.765	0.748	0.534	0.404	0.364	0.414
-2.8	0.358	0.258	0.208	0.208	0.688	0.541	0.623	0.582	0.358	0.258	0.208	0.208
-2.6	0.261	0.171	0.111	0.061	0.575	0.457	0.516	0.483	0.261	0.171	0.111	0.061
-2.4	0.241	0.142	0.041	0.06	0.449	0.425	0.391	0.464	0.241	0.142	0.041	0.06
-2.2	0.161	0.081	-0.059	-0.039	0.376	0.359	0.348	0.368	0.161	0.081	0.059	0.039
-2	0.102	0.042	-0.098	-0.088	0.34	0.292	0.301	0.327	0.102	0.042	0.098	0.088
-1.8	0.071	0.001	-0.149	-0.119	0.321	0.331	0.338	0.365	0.071	0.001	0.149	0.119
-1.6	0.054	-0.036	-0.146	-0.126	0.318	0.345	0.388	0.385	0.054	0.036	0.146	0.126
-1.4	0.008	-0.092	-0.192	-0.162	0.264	0.343	0.395	0.416	0.008	0.092	0.192	0.162
-1.2	-0.039	-0.149	-0.249	-0.219	0.251	0.337	0.373	0.399	0.039	0.149	0.249	0.219
-1	0.005	-0.107	-0.187	-0.18	0.284	0.332	0.396	0.409	0.005	0.107	0.187	0.18
-0.8	0.003	-0.144	-0.224	-0.224	0.267	0.349	0.387	0.427	0.003	0.144	0.224	0.224
-0.6	-0.003	-0.174	-0.264	-0.294	0.266	0.313	0.373	0.437	0.003	0.174	0.264	0.294
-0.4	-0.02	-0.155	-0.255	-0.285	0.281	0.311	0.36	0.399	0.02	0.155	0.255	0.285
-0.2	-0.022	-0.153	-0.252	-0.281	0.27	0.357	0.386	0.439	0.022	0.153	0.252	0.281
0	-0.035	-0.171	-0.272	-0.323	0.252	0.35	0.429	0.456	0.035	0.171	0.272	0.323
0.2	0.005	-0.197	-0.267	-0.337	0.249	0.375	0.467	0.482	0.005	0.197	0.267	0.337
0.4	0.003	-0.201	-0.261	-0.311	0.2556	0.383	0.481	0.52	0.003	0.201	0.261	0.311
0.6	-0.003	-0.192	-0.232	-0.292	0.267	0.417	0.478	0.51	0.003	0.192	0.232	0.292
0.8	-0.02	-0.184	-0.194	-0.254	0.243	0.418	0.418	0.49	0.02	0.184	0.194	0.254
1	-0.035	-0.227	-0.247	-0.267	0.28	0.393	0.422	0.462	0.035	0.227	0.247	0.267
1.2	-0.142	-0.326	-0.326	-0.346	0.287	0.402	0.401	0.455	0.142	0.326	0.326	0.346
1.4	-0.211	-0.366	-0.386	-0.386	0.286	0.342	0.387	0.339	0.211	0.366	0.386	0.386
1.6	-0.341	-0.384	-0.412	-0.401	0.337	0.338	0.335	0.389	0.341	0.384	0.412	0.401
1.8	-0.393	-0.42	-0.413	-0.432	0.384	0.377	0.429	0.453	0.393	0.42	0.413	0.433
2	-0.469	-0.459	-0.479	-0.479	0.402	0.442	0.428	0.436	0.469	0.459	0.479	0.479
2.2	-0.531	-0.581	-0.611	-0.641	0.465	0.496	0.502	0.532	0.531	0.581	0.611	0.641
2.4	-0.633	-0.733	-0.683	-0.753	0.558	0.521	0.591	0.571	0.633	0.733	0.683	0.753
2.6	-0.701	-0.721	-0.781	-0.821	0.619	0.634	0.602	0.673	0.701	0.721	0.781	0.821
2.8	-0.826	-0.876	-0.906	-0.856	0.667	0.706	0.776	0.706	0.826	0.876	0.906	0.856
3	-0.912	-0.963	-0.91	-0.861	0.733	0.806	0.868	0.784	0.912	0.963	0.91	0.861

Table 4. 9 Evaluation Indices of Ability Score Recovery Conditioned on Two-Stage Panel Configuration (1-3)

Theta	Bias			RMSE			MAE		
	SC	SP1	SP2	SC	SP1	SP2	SC	SP1	SP2
-3	0.661	0.567	0.587	0.891	0.72	0.769	0.661	0.567	0.587
-2.8	0.506	0.423	0.305	0.723	0.56	0.599	0.506	0.423	0.305
-2.6	0.401	0.315	0.193	0.632	0.492	0.513	0.401	0.315	0.193
-2.4	0.312	0.237	0.087	0.542	0.429	0.401	0.312	0.237	0.087
-2.2	0.254	0.154	0.028	0.433	0.375	0.321	0.254	0.154	0.028
-2	0.201	0.055	-0.076	0.453	0.339	0.301	0.201	0.055	0.076
-1.8	0.165	0.024	-0.095	0.376	0.358	0.329	0.165	0.024	0.095
-1.6	0.122	-0.022	-0.146	0.351	0.398	0.399	0.122	0.022	0.146
-1.4	0.085	-0.044	-0.151	0.333	0.368	0.419	0.085	0.044	0.151
-1.2	0.042	-0.041	-0.159	0.312	0.34	0.389	0.042	0.041	0.159
-1	0.022	-0.063	-0.175	0.346	0.362	0.401	0.022	0.063	0.175
-0.8	0.003	-0.088	-0.184	0.364	0.349	0.37	0.003	0.088	0.184
-0.6	0.005	-0.08	-0.221	0.292	0.333	0.372	0.005	0.08	0.221
-0.4	0.003	-0.088	-0.209	0.372	0.337	0.363	0.003	0.088	0.209
-0.2	-0.003	-0.087	-0.223	0.313	0.357	0.395	0.003	0.087	0.223
0	-0.02	-0.115	-0.228	0.292	0.375	0.421	0.02	0.115	0.228
0.2	-0.022	-0.112	-0.204	0.296	0.415	0.449	0.022	0.112	0.204
0.4	-0.035	-0.125	-0.222	0.321	0.438	0.483	0.035	0.125	0.222
0.6	-0.038	-0.117	-0.25	0.295	0.459	0.481	0.038	0.117	0.25
0.8	-0.064	-0.143	-0.287	0.292	0.441	0.438	0.064	0.143	0.287
1	-0.093	-0.178	-0.278	0.316	0.339	0.312	0.093	0.178	0.278
1.2	-0.127	-0.152	-0.252	0.317	0.352	0.38	0.127	0.152	0.252
1.4	-0.159	-0.164	-0.194	0.334	0.362	0.388	0.159	0.164	0.194
1.6	-0.183	-0.184	-0.258	0.376	0.389	0.425	0.183	0.184	0.258
1.8	-0.251	-0.282	-0.229	0.411	0.412	0.44	0.251	0.282	0.229
2	-0.326	-0.318	-0.284	0.452	0.492	0.514	0.326	0.318	0.284
2.2	-0.387	-0.408	-0.415	0.536	0.532	0.557	0.387	0.408	0.415
2.4	-0.48	-0.522	-0.506	0.587	0.558	0.585	0.48	0.522	0.506
2.6	-0.591	-0.627	-0.701	0.642	0.672	0.703	0.591	0.627	0.701
2.8	-0.68	-0.646	-0.667	0.716	0.754	0.782	0.68	0.646	0.667
3	-0.784	-0.82	-0.873	0.824	0.849	0.881	0.784	0.82	0.873

Table 4. 10 Evaluation Indices of Ability Score Recovery Conditioned on Three-Stage Panel Configuration (1-2-3)

Theta	Bias				RMSE				MAE			
	SC	SP1	SP2	SP3	SC	SP1	SP2	SP3	SC	SP1	SP2	SP3
-3	0.503	0.371	0.388	0.37	0.86	0.674	0.769	0.734	0.503	0.371	0.388	0.37
-2.8	0.402	0.308	0.231	0.202	0.694	0.532	0.599	0.592	0.402	0.308	0.231	0.202
-2.6	0.289	0.187	0.12	0.072	0.556	0.454	0.513	0.494	0.289	0.187	0.12	0.072
-2.4	0.222	0.105	0.053	0.021	0.46	0.404	0.401	0.455	0.222	0.105	0.053	0.021
-2.2	0.193	0.079	0.015	-0.004	0.384	0.338	0.321	0.368	0.193	0.079	0.015	0.004
-2	0.155	0.058	0.002	0.002	0.362	0.293	0.301	0.335	0.155	0.058	0.002	0.002
-1.8	0.102	0.006	-0.05	-0.092	0.329	0.333	0.329	0.37	0.102	0.006	0.05	0.092
-1.6	0.063	0.002	-0.122	-0.152	0.313	0.354	0.399	0.391	0.063	0.002	0.122	0.152
-1.4	0.032	-0.021	-0.154	-0.179	0.288	0.348	0.419	0.399	0.032	0.021	0.154	0.179
-1.2	0.011	-0.055	-0.157	-0.193	0.262	0.349	0.389	0.398	0.011	0.055	0.157	0.193
-1	0.006	-0.092	-0.185	-0.185	0.265	0.324	0.401	0.379	0.006	0.092	0.185	0.185
-0.8	0.003	-0.085	-0.167	-0.172	0.275	0.329	0.37	0.43	0.003	0.085	0.167	0.172
-0.6	0.005	-0.112	-0.203	-0.246	0.242	0.31	0.372	0.432	0.005	0.112	0.203	0.246
-0.4	0.003	-0.095	-0.227	-0.238	0.259	0.312	0.363	0.403	0.003	0.095	0.227	0.238
-0.2	-0.003	-0.13	-0.256	-0.241	0.273	0.333	0.395	0.446	0.003	0.13	0.256	0.241
0	-0.023	-0.156	-0.24	-0.229	0.256	0.353	0.421	0.451	0.02	0.156	0.24	0.229
0.2	-0.022	-0.164	-0.249	-0.284	0.262	0.394	0.449	0.491	0.022	0.164	0.249	0.284
0.4	-0.035	-0.156	-0.251	-0.324	0.27	0.409	0.483	0.52	0.035	0.156	0.251	0.324
0.6	-0.038	-0.14	-0.245	-0.312	0.266	0.415	0.481	0.518	0.038	0.14	0.245	0.312
0.8	-0.064	-0.153	-0.258	-0.325	0.259	0.418	0.438	0.489	0.064	0.153	0.258	0.325
1	-0.093	-0.178	-0.253	-0.312	0.281	0.398	0.456	0.472	0.093	0.178	0.253	0.312
1.2	-0.117	-0.151	-0.176	-0.25	0.283	0.382	0.425	0.469	0.11	0.151	0.176	0.25
1.4	-0.109	-0.133	-0.172	-0.242	0.299	0.341	0.352	0.332	0.109	0.133	0.172	0.242
1.6	-0.167	-0.155	-0.182	-0.201	0.324	0.357	0.375	0.384	0.167	0.155	0.182	0.201
1.8	-0.226	-0.242	-0.22	-0.267	0.376	0.389	0.44	0.436	0.226	0.242	0.22	0.267
2	-0.28	-0.253	-0.278	-0.347	0.419	0.444	0.514	0.459	0.28	0.253	0.278	0.347
2.2	-0.355	-0.379	-0.401	-0.466	0.481	0.504	0.557	0.532	0.355	0.379	0.401	0.466
2.4	-0.455	-0.483	-0.57	-0.552	0.544	0.516	0.585	0.574	0.455	0.483	0.57	0.552
2.6	-0.572	-0.512	-0.58	-0.608	0.615	0.64	0.703	0.657	0.572	0.512	0.58	0.608
2.8	-0.643	-0.721	-0.732	-0.792	0.674	0.709	0.782	0.711	0.643	0.721	0.732	0.792
3	-0.765	-0.795	-0.814	-0.791	0.752	0.817	0.881	0.789	0.765	0.795	0.814	0.791

Table 4. 11 Evaluation Indices of Ability Score Recovery Conditioned on Three-Stage Panel Configuration (1-3-3)

Theta	Bias				RMSE				MAE			
	SC	SP1	SP2	SP3	SC	SP1	SP2	SP3	SC	SP1	SP2	SP3
-3	0.491	0.453	0.38	0.302	0.812	0.696	0.792	0.77	0.491	0.453	0.38	0.302
-2.8	0.377	0.293	0.205	0.125	0.649	0.558	0.624	0.6	0.377	0.293	0.205	0.125
-2.6	0.272	0.164	0.112	0.024	0.507	0.482	0.538	0.55	0.272	0.164	0.112	0.024
-2.4	0.207	0.079	-0.003	-0.0253	0.415	0.431	0.424	0.511	0.207	0.079	0.003	0.0253
-2.2	0.166	0.059	-0.023	-0.055	0.326	0.362	0.345	0.438	0.166	0.059	0.023	0.055
-2	0.138	0.031	-0.045	-0.089	0.332	0.317	0.336	0.363	0.138	0.031	0.045	0.089
-1.8	0.091	-0.022	-0.133	-0.203	0.285	0.36	0.36	0.424	0.091	0.022	0.133	0.203
-1.6	0.041	-0.031	-0.172	-0.224	0.263	0.384	0.423	0.45	0.041	0.031	0.172	0.224
-1.4	0.008	-0.051	-0.167	-0.239	0.257	0.368	0.453	0.44	0.008	0.051	0.167	0.239
-1.2	-0.004	-0.074	-0.223	-0.247	0.213	0.387	0.422	0.463	0.004	0.074	0.223	0.247
-1	0.006	-0.116	-0.219	-0.255	0.232	0.356	0.432	0.472	0.006	0.116	0.219	0.255
-0.8	0.003	-0.114	-0.218	-0.255	0.236	0.366	0.404	0.477	0.003	0.114	0.218	0.255
-0.6	0.005	-0.122	-0.232	-0.251	0.209	0.331	0.402	0.443	0.005	0.122	0.232	0.251
-0.4	0.003	-0.123	-0.241	-0.251	0.202	0.343	0.386	0.457	0.003	0.123	0.241	0.251
-0.2	-0.003	-0.152	-0.249	-0.288	0.217	0.372	0.428	0.488	0.003	0.152	0.249	0.288
0	-0.021	-0.169	-0.278	-0.29	0.226	0.38	0.46	0.536	0.021	0.169	0.278	0.29
0.2	-0.022	-0.179	-0.296	-0.32	0.218	0.419	0.482	0.535	0.022	0.179	0.296	0.32
0.4	-0.035	-0.178	-0.258	-0.312	0.217	0.443	0.516	0.579	0.035	0.178	0.258	0.312
0.6	-0.038	-0.155	-0.231	-0.335	0.222	0.448	0.502	0.565	0.038	0.155	0.231	0.335
0.8	-0.064	-0.143	-0.239	-0.302	0.209	0.449	0.472	0.528	0.064	0.143	0.239	0.302
1	-0.082	-0.138	-0.206	-0.298	0.228	0.438	0.442	0.533	0.082	0.138	0.206	0.298
1.2	-0.112	-0.135	-0.186	-0.281	0.221	0.402	0.451	0.492	0.112	0.135	0.186	0.281
1.4	-0.136	-0.141	-0.169	-0.203	0.269	0.289	0.309	0.314	0.136	0.141	0.169	0.203
1.6	-0.182	-0.178	-0.218	-0.188	0.293	0.328	0.325	0.337	0.182	0.178	0.218	0.188
1.8	-0.247	-0.261	-0.31	-0.283	0.327	0.35	0.357	0.388	0.247	0.261	0.31	0.283
2	-0.303	-0.272	-0.304	-0.274	0.368	0.408	0.4	0.422	0.303	0.272	0.304	0.274
2.2	-0.371	-0.403	-0.433	-0.404	0.447	0.472	0.479	0.442	0.371	0.403	0.433	0.404
2.4	-0.474	-0.494	-0.516	-0.486	0.489	0.52	0.512	0.565	0.474	0.494	0.516	0.486
2.6	-0.585	-0.524	-0.561	-0.547	0.581	0.612	0.607	0.642	0.585	0.524	0.561	0.547
2.8	-0.658	-0.733	-0.763	-0.726	0.63	0.654	0.65	0.624	0.658	0.733	0.763	0.726
3	-0.785	-0.815	-0.827	-0.807	0.701	0.73	0.726	0.771	0.785	0.815	0.827	0.807

Table 4. 12 Evaluation Indices of Ability Score Recovery Conditioned on Low Item Pool Difficulty

Theta	Bias				RMSE				MAE			
	SC	SP1	SP2	SP3	SC	SP1	SP2	SP3	SC	SP1	SP2	SP3
-3	0.289	0.266	0.245	0.249	0.602	0.623	0.662	0.655	0.289	0.266	0.245	0.249
-2.8	0.221	0.189	0.163	0.184	0.531	0.525	0.56	0.546	0.221	0.189	0.163	0.184
-2.6	0.183	0.144	0.105	0.135	0.423	0.449	0.464	0.469	0.183	0.144	0.105	0.135
-2.4	0.135	0.108	0.068	0.083	0.385	0.392	0.422	0.419	0.135	0.108	0.068	0.08
-2.2	0.086	0.047	0.015	0.036	0.356	0.368	0.407	0.398	0.086	0.047	0.015	0.036
-2	0.042	0.013	-0.024	0.004	0.332	0.347	0.367	0.381	0.042	0.013	0.024	0.004
-1.8	0.023	-0.014	-0.052	-0.027	0.314	0.336	0.373	0.362	0.023	0.014	0.052	0.027
-1.6	0.014	-0.019	-0.057	-0.037	0.313	0.344	0.355	0.377	0.014	0.019	0.057	0.037
-1.4	0.006	-0.034	-0.063	-0.035	0.301	0.328	0.346	0.368	0.006	0.034	0.063	0.035
-1.2	0.003	-0.019	-0.053	-0.028	0.304	0.328	0.341	0.362	0.003	0.019	0.053	0.028
-1	0.005	-0.022	-0.052	-0.043	0.309	0.325	0.344	0.352	0.005	0.022	0.052	0.043
-0.8	0.003	-0.018	-0.055	-0.034	0.289	0.323	0.335	0.344	0.003	0.018	0.055	0.034
-0.6	-0.003	-0.037	-0.068	-0.044	0.282	0.299	0.32	0.332	0.003	0.037	0.068	0.044
-0.4	-0.02	-0.051	-0.085	-0.066	0.302	0.328	0.341	0.361	0.02	0.051	0.085	0.066
-0.2	-0.022	-0.05	-0.078	-0.055	0.293	0.305	0.337	0.332	0.022	0.05	0.078	0.055
0	-0.035	-0.058	-0.096	-0.072	0.287	0.306	0.326	0.336	0.035	0.058	0.096	0.072
0.2	-0.038	-0.073	-0.112	-0.085	0.312	0.325	0.351	0.365	0.038	0.073	0.112	0.085
0.4	-0.064	-0.103	-0.124	-0.096	0.317	0.333	0.345	0.369	0.064	0.103	0.124	0.096
0.6	-0.093	-0.103	-0.121	-0.128	0.332	0.347	0.378	0.386	0.093	0.103	0.121	0.128
0.8	-0.112	-0.123	-0.141	-0.147	0.359	0.371	0.361	0.381	0.112	0.123	0.141	0.147
1	-0.119	-0.138	-0.153	-0.138	0.319	0.33	0.356	0.371	0.119	0.138	0.153	0.138
1.2	-0.167	-0.184	-0.202	-0.181	0.332	0.35	0.372	0.374	0.167	0.184	0.202	0.181
1.4	-0.202	-0.213	-0.231	-0.213	0.352	0.372	0.404	0.395	0.202	0.213	0.231	0.213
1.6	-0.25	-0.263	-0.273	-0.264	0.378	0.411	0.449	0.448	0.25	0.263	0.273	0.264
1.8	-0.302	-0.312	-0.329	-0.31	0.435	0.47	0.492	0.504	0.302	0.312	0.329	0.31
2	-0.382	-0.392	-0.405	-0.391	0.503	0.537	0.564	0.561	0.382	0.392	0.405	0.391
2.2	-0.47	-0.482	-0.495	-0.487	0.583	0.605	0.643	0.636	0.47	0.482	0.495	0.487
2.4	-0.553	-0.565	-0.585	-0.566	0.68	0.706	0.723	0.729	0.553	0.565	0.585	0.566
2.6	-0.62	-0.632	-0.648	-0.636	0.792	0.811	0.848	0.833	0.62	0.632	0.648	0.636
2.8	-0.702	-0.711	-0.721	-0.713	0.879	0.909	0.941	0.937	0.702	0.711	0.721	0.713
3	-0.751	-0.765	-0.784	-0.766	1.03	1.051	1.076	1.091	0.751	0.765	0.784	0.766

Table 4. 13 Evaluation Indices of Ability Score Recovery Conditioned on Moderate Item Pool Difficulty

Theta	Bias				RMSE				MAE			
	SC	SP1	SP2	SP3	SC	SP1	SP2	SP3	SC	SP1	SP2	SP3
-3	0.593	0.403	0.429	0.409	0.841	0.643	0.741	0.706	0.593	0.403	0.429	0.409
-2.8	0.432	0.341	0.26	0.222	0.668	0.483	0.578	0.571	0.432	0.341	0.26	0.222
-2.6	0.332	0.22	0.114	0.083	0.534	0.402	0.489	0.47	0.332	0.22	0.114	0.083
-2.4	0.263	0.153	0.084	0.053	0.417	0.357	0.37	0.424	0.263	0.153	0.084	0.053
-2.2	0.228	0.128	0.048	0.03	0.369	0.301	0.304	0.351	0.228	0.128	0.048	0.03
-2	0.189	0.108	0.016	0.002	0.333	0.261	0.253	0.287	0.189	0.108	0.016	0.002
-1.8	0.121	0.051	-0.016	-0.048	0.307	0.291	0.295	0.336	0.121	0.051	0.016	0.048
-1.6	0.093	0.013	-0.092	-0.109	0.285	0.32	0.366	0.358	0.093	0.013	0.092	0.109
-1.4	0.053	-0.007	-0.107	-0.162	0.269	0.298	0.399	0.379	0.053	0.007	0.107	0.162
-1.2	0.052	-0.01	-0.116	-0.18	0.222	0.321	0.37	0.379	0.052	0.01	0.116	0.18
-1	0.052	-0.073	-0.162	-0.151	0.216	0.282	0.366	0.368	0.052	0.073	0.162	0.151
-0.8	0.018	-0.052	-0.153	-0.201	0.225	0.296	0.351	0.411	0.018	0.052	0.153	0.201
-0.6	0.037	-0.093	-0.185	-0.215	0.202	0.302	0.351	0.411	0.037	0.093	0.185	0.215
-0.4	0.014	-0.07	-0.207	-0.194	0.211	0.295	0.328	0.386	0.014	0.07	0.207	0.194
-0.2	0.033	-0.105	-0.223	-0.207	0.222	0.321	0.367	0.418	0.033	0.105	0.223	0.207
0	0.029	-0.134	-0.203	-0.213	0.209	0.32	0.383	0.413	0.029	0.134	0.203	0.213
0.2	-0.008	-0.127	-0.235	-0.248	0.218	0.348	0.416	0.458	0.008	0.127	0.235	0.248
0.4	0.005	-0.131	-0.232	-0.275	0.222	0.388	0.471	0.508	0.005	0.131	0.232	0.275
0.6	-0.022	-0.101	-0.226	-0.279	0.233	0.393	0.442	0.479	0.022	0.101	0.226	0.279
0.8	-0.049	-0.126	-0.24	-0.289	0.224	0.391	0.418	0.469	0.049	0.126	0.24	0.289
1	-0.063	-0.103	-0.234	-0.301	0.259	0.271	0.302	0.295	0.063	0.103	0.234	0.301
1.2	-0.066	-0.116	-0.151	-0.224	0.262	0.292	0.325	0.289	0.066	0.116	0.151	0.224
1.4	-0.063	-0.105	-0.142	-0.206	0.284	0.335	0.312	0.321	0.063	0.105	0.142	0.206
1.6	-0.101	-0.113	-0.146	-0.19	0.292	0.322	0.387	0.37	0.101	0.113	0.146	0.19
1.8	-0.191	-0.229	-0.195	-0.221	0.34	0.359	0.404	0.4	0.191	0.229	0.195	0.221
2	-0.264	-0.236	-0.265	-0.32	0.388	0.396	0.466	0.411	0.264	0.236	0.265	0.32
2.2	-0.318	-0.347	-0.385	-0.451	0.433	0.469	0.492	0.498	0.318	0.347	0.385	0.451
2.4	-0.437	-0.447	-0.546	-0.521	0.517	0.496	0.554	0.543	0.437	0.447	0.546	0.521
2.6	-0.52	-0.462	-0.533	-0.572	0.598	0.602	0.671	0.625	0.52	0.462	0.533	0.572
2.8	-0.596	-0.705	-0.705	-0.747	0.624	0.624	0.712	0.7	0.596	0.705	0.705	0.747
3	-0.746	-0.771	-0.791	-0.759	0.708	0.785	0.793	0.746	0.746	0.771	0.791	0.759

Table 4. 14 Evaluation Indices of Ability Score Recovery Conditioned on High Item Pool Difficulty

Theta	Bias				RMSE				MAE			
	SC	SP1	SP2	SP3	SC	SP1	SP2	SP3	SC	SP1	SP2	SP3
-3	0.933	0.847	0.652	0.59	1.332	1.15	1.031	1.013	0.933	0.847	0.652	0.59
-2.8	0.852	0.735	0.636	0.492	1.165	1.035	0.935	0.852	0.852	0.735	0.636	0.492
-2.6	0.811	0.72	0.552	0.401	1.032	0.958	0.823	0.765	0.811	0.72	0.552	0.401
-2.4	0.755	0.667	0.462	0.324	0.946	0.869	0.752	0.633	0.755	0.667	0.462	0.324
-2.2	0.672	0.555	0.402	0.264	0.921	0.841	0.716	0.526	0.672	0.555	0.402	0.264
-2	0.601	0.495	0.324	0.167	0.842	0.77	0.573	0.392	0.601	0.495	0.324	0.167
-1.8	0.522	0.409	0.25	0.099	0.833	0.692	0.495	0.315	0.522	0.409	0.25	0.099
-1.6	0.445	0.328	0.192	0.055	0.767	0.652	0.432	0.303	0.445	0.328	0.192	0.055
-1.4	0.367	0.244	0.105	-0.052	0.714	0.625	0.382	0.325	0.367	0.244	0.105	0.052
-1.2	0.292	0.182	0.093	-0.051	0.626	0.562	0.351	0.343	0.292	0.182	0.093	0.051
-1	0.223	0.145	0.056	-0.106	0.593	0.538	0.311	0.393	0.223	0.145	0.056	0.106
-0.8	0.184	0.094	0.032	-0.112	0.534	0.471	0.35	0.426	0.184	0.094	0.032	0.112
-0.6	0.132	0.057	0.002	-0.123	0.488	0.427	0.343	0.471	0.132	0.057	0.002	0.123
-0.4	0.085	0.015	-0.063	-0.207	0.442	0.384	0.356	0.523	0.085	0.015	0.063	0.207
-0.2	0.061	-0.022	-0.068	-0.183	0.431	0.343	0.421	0.514	0.061	0.022	0.068	0.183
0	0.042	-0.031	-0.077	-0.223	0.362	0.313	0.413	0.503	0.042	0.031	0.077	0.223
0.2	0.026	-0.069	-0.121	-0.269	0.343	0.302	0.453	0.572	0.026	0.069	0.121	0.269
0.4	0.017	-0.065	-0.13	-0.271	0.322	0.355	0.491	0.596	0.017	0.065	0.13	0.271
0.6	0.006	-0.047	-0.128	-0.252	0.311	0.371	0.515	0.613	0.006	0.047	0.128	0.252
0.8	0.003	-0.061	-0.14	-0.243	0.302	0.412	0.532	0.637	0.003	0.061	0.14	0.243
1	0.005	-0.057	-0.134	-0.254	0.303	0.423	0.525	0.646	0.005	0.057	0.134	0.254
1.2	0.003	-0.081	-0.126	-0.249	0.292	0.444	0.557	0.675	0.003	0.081	0.126	0.249
1.4	-0.003	-0.058	-0.135	-0.231	0.312	0.452	0.539	0.637	0.003	0.058	0.135	0.231
1.6	-0.022	-0.063	-0.084	-0.201	0.318	0.453	0.523	0.596	0.022	0.063	0.084	0.201
1.8	-0.022	-0.046	-0.056	-0.072	0.333	0.351	0.363	0.387	0.022	0.046	0.056	0.072
2	-0.05	-0.064	-0.067	-0.096	0.366	0.403	0.384	0.336	0.05	0.064	0.067	0.096
2.2	-0.132	-0.13	-0.113	-0.16	0.432	0.421	0.433	0.455	0.132	0.13	0.113	0.16
2.4	-0.135	-0.145	-0.157	-0.167	0.483	0.522	0.52	0.534	0.135	0.145	0.157	0.167
2.6	-0.192	-0.203	-0.229	-0.213	0.525	0.539	0.554	0.519	0.192	0.203	0.229	0.213
2.8	-0.253	-0.276	-0.284	-0.305	0.622	0.653	0.656	0.692	0.253	0.276	0.284	0.305
3	-0.354	-0.384	-0.39	-0.404	0.712	0.742	0.738	0.704	0.354	0.384	0.39	0.404