EXAMINING CONTENT CONTROL IN ADAPTIVE TESTS: COMPUTERIZED ADAPTIVE TESTING VS. COMPUTERIZED MULTISTAGE TESTING

By

HALIL IBRAHIM SARI

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

2016

© 2016 Halil Ibrahim Sari

To my family

ACKNOWLEDGMENTS

First of all, I would like to thank my patient advisor Dr. Anne Corinne Huggins-Manley for guiding me in my dissertation. I would not have completed this dissertation without her help. I am proud to be her graduating advisee. I would also like to thank to my committee members Drs. David Miller, Tim Jacobbe and Tamer Kahveci for their valuable feedback on earlier versions of this dissertation.

Secondly, I would like to thank the Turkish Government for providing a full scholarship to pursue my graduate studies in the U.S. Without this scholarship, I would not be able to afford my graduate studies. I also thank Necla Sari and Zeki Uyar for trusting me.

Lastly and most importantly, I am thankful to my beloved wife, Hasibe Yahsi Sari and my adorable daughter, Erva Bilge Sari, who have always been there as a motivational source for me. I would not be able to complete this dissertation if they did not show understanding regarding my abnormal work schedule. I am also thankful in advance to my baby girl that I am expecting because I am getting rid of dissertation stress early thanks to her.

TABLE OF CONTENTS

page

ACKNOWLEDGMENTS 4

LIST OF TABLES 7

LIST OF FIGURES 8

LIST OF ABBREVIATIONS 9

ABSTRACT 10

CHAPTER

1 INTRODUCTION 12

Test Administration Models 12
Connection of Content Control with Test Fairness and Validity 14

2 LITERATURE REVIEW 18

Item Response Theory 18
Assumptions of Item Response Theory 19
Dichotomous Item Response Theory Models 20
Item and Test Information in IRT 21
Overview of Computerized Adaptive Testing 23
Item Bank 25
Item Selection Rule 26
Ability Estimation 28
Item Exposure Control 30
Content Control 32
Stopping Rule 34
Overview of Computerized Multistage Testing 35
Test Design and Structure 39
Number of panels 39
Number stages and modules 39
Number of items 40
Routing Method 41
Automated Test Assembly, Content Control and Exposure Control 43
Statement of the Problem 44

3 METHODOLOGY 53

Design Overview 53
Fixed Conditions 53
CAT Simulations 55

Ca-MST Simulations 57
Evaluation Criteria 59

4 RESULTS 67

Overall Results 67
Mean Bias 67
Root Mean Square Error 68
Correlation 68
Conditional 69

5 DISCUSSION AND LIMITATIONS 80

LIST OF REFERENCES 86

BIOGRAPHICAL SKETCH 95

LIST OF TABLES

Table page

1-1 Main advantages and disadvantages of test administration models.....17

3-1 Item parameters of each content area in the original item bank 61

3-2 Item parameters of each content area in item bank 1.....62

3-3 Distribution of items across the contents areas and test lengths in CAT.....63

3-4 The distribution of items in modules in ca-MST across the content areas under 24-item test length.....64

3-5 The distribution of items in modules in ca-MSTs across the content areas in under 48-item test length.....65

4-1 Results of mean bias across the conditions.....70

4-2 Factorial ANOVA findings when dependent variable was mean bias 71

4-3 Results of RMSE across the conditions.....72

4-4 Factorial ANOVA findings when dependent variable was RMSE.....73

4-5 Correlation coefficients between true and estimated thetas across the conditions 74

4-6 Factorial ANOVA findings when dependent variable was correlation 75

LIST OF FIGURES

Figure page

2-1 An example of item characteristic curve (ICC). 47

2-2 Total test information functions for three different tests. 48

2-3 The flowchart of computerized adaptive testing.....49

2-4 An example of 1-2 ca-MST panel design. 50

2-5 An example of 1-3-3 ca-MST panel design 51

2-6 Illustration of multiple panels in ca-MST. 52

3-1 Total information functions for four item banks.....66

4-1 Main effect of test administration model on mean bias within levels of test length. 76

4-2 Interaction of test administration model and test length on RMSE within levels of number of content. 77

4-3 Main effect of test administration model on correlation within levels of test length. 78

4-4 Conditional standard error of measurements 79

LIST OF ABBREVIATIONS

Ca-MST	Computerized Multistage Testing
CAT	Computerized Adaptive Testing
CSEM	Conditional Standard Error of Measurement
CTT	Classical Test Theory
ICC	Item Characteristic Curve
IRT	Item Response Theory
RMSE	Root Mean Square Error
SEM	Standard Error of Measurement

Abstract of Dissertation Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy

EXAMINING CONTENT CONTROL IN ADAPTIVE TESTS:
COMPUTERIZED ADAPTIVE TESTING VS. COMPUTERIZED MULTISTAGE TESTING

By

Halil Ibrahim Sari

August 2016

Chair: Anne Corinne Huggins-Manley
Major: Research and Evaluation Methodology

Many comparison studies have been conducted to investigate efficiency of the statistical procedures across computerized adaptive testing (CAT) and computerized multistage testing (ca-MST). Although it is directly related to validity, score interpretation and test fairness, non-statistical issues of adaptive tests, such as content balancing, have not been given more attention. It is consistently asserted in several studies that major advantage of ca-MST is that it controls for content better than CAT. Yet, the literature does not contain a study that specifically compares CAT with ca-MST under varying levels of content constraints to verify this claim. A simulation study was conducted to explore the precision of test outcomes across CAT and ca-MST when the number of different content areas was varied across a variety of test lengths. One CAT and two ca-MST designs (1-3 and 1-3-3 panel designs) were compared across several manipulated conditions including total test length (24-item and 48-item test length) and number of controlled content areas. The five levels of the content area condition included zero (no content control), two, four, six and eight content area. All manipulated conditions within CAT and ca-MST were fully crossed with one another. This resulted in 2x5=10 CAT (test length x content area), and 2x5x2=20 ca-MST conditions (test length x content area x ca-MST panel design), for 30 total

conditions. 4000 examinees were generated from $N(0,1)$. All other conditions such as IRT model, exposure rate were fixed across the CAT and ca-MSTs. Results were evaluated with mean bias, root mean square error, correlation between true and estimated thetas and conditional standard error of measurement. Results illustrated that test length and the type of test administration model impacted the outcomes more than the number of content area. The main finding was that regardless of any study condition, CAT outperformed the two ca-MSTs, and the two ca-MSTs were comparable. The results were discussed in connection to the control over test design, test content, cost effectiveness and item pool usage. Recommendations for practitioner were provided. Limitations for further research were listed.

CHAPTER 1
INTRODUCTION

Test Administration Models

The increased role of computers in educational and psychological measurement has led testing companies (e.g., Educational Testing Service, Pearson, KAPLAN, The College Board, and American College Testing) and practitioners to explore alternative ways to measure student achievement. There are numerous test administration models. Each model has some pros and cons in terms of test validity, score reliability, test fairness and cost. The most widely used traditional model today is the linear form. In this testing approach, the exam is administered on paper, the same set of items are given to all examinees, and item order cannot change during the test (e.g., American College Testing-ACT). Since all examinees receive the same set of items, it is relatively easy to construct the forms because it does not require an item pool, which requires additional time, effort and money. One of the big advantages of linear models is high test developer control on content. This means that prior to the test administration, for each subject (e.g., biology) practitioners can specify related content areas (e.g., photosynthesis, ecology, plants, human anatomy, animals) and the number of items needed within each content area. Consistent content among all examinees is a vital issue because it is strongly related to test fairness (Wainer, Dorans, Flaugher, Green, & Mislevy, 2000) and validity (Yan, von Davier, & Lewis, 2014). For example, on a biology test if a test taker who is strong in ecology but struggles in other areas receives many ecology items but receives only a few items from other areas, this will most likely cause overestimation of her biology score. In this case, test validity will be threatened (Leung, Chang, & Hau, 2002).

In addition, easier test administration, flexible scheduling, flexible item format, and item review by the examinees are primary advantages associated with linear tests (Becker &

Bergstrom, 2013). However, one big criticism of linear tests is the inevitable jeopardizing of test security (e.g., cheating). This is because each test item is exposed to all test takers, which is a serious threat for test validity and score reliability (Thompson, 2008). It is possible to see some examples of linear tests that overcome item exposure rate and test security by creating multiple linear forms (e.g., SAT). But, drawback to this testing model is delayed scoring and late reporting. This requires test takers to wait for their test scores to become available, which can present a problem if test takers are near application deadlines relatively. Long test length and low measurement efficiency can be counted as other major disadvantages of linear tests (Yan et al., 2014).

Linear tests can also be administrated on computers (e.g., linear computer based test or computer fixed tests). In this type of testing approach, the exam is delivered via computers. Linear CBTs can overcome some of the deficiencies of linear tests such as delayed reporting and late scoring but fail to deal with other deficiencies such as low measurement accuracy and long test length (Becker & Bergstrom, 2013).

Another form of the test administration model is adaptive testing. In this approach, as in the linear computer based tests, the exam is delivered on a computer, and the computer algorithms determine which items the examinee will receive based on her performance. There are two versions of adaptive testing; computerized adaptive testing (CAT) and computerized multistage testing (ca-MST). In CAT (Weiss & Kingsbury, 1984), the test taker starts the exam with an item that is at medium difficulty and then, depending on the performance on this item, the computer algorithm selects the next question that contributes the most information about her current ability from the item pool. This means that if a test taker gets the first item correct, the computer administers a more difficult item, if wrong, an easier item. In ca-MST, the test taker

receives a set of items instead individual items. She starts the exam with a testlet (e.g., a set of 5 or 10 items). Depending on the performance on this testlet, the computer selects the next testlet that contributes the highest information about the test taker's current ability. This means that if she shows poor performance on the first testlet, she receives an easier testlet, otherwise she receives a more difficult testlet. As implied, the difference between ca-MST and CAT is that there is a testlet level adaptation in ca-MST, so the number of adaptation points is less than CAT. Both CAT and ca-MST have multiple advantages over other administration models, such as immediate scoring, high measurement accuracy, low test length and high test security (Yan et al., 2014). However, one drawback to adaptive testing models is that it is not very easy to ensure all examinees are exposed to the same distribution of items in terms of content. This is extremely critical, especially in high stakes administrations (Huang, 1996), because unless the content distribution of items is the same across examinees, the test is essentially measuring different constructs for different persons. The main advantages and disadvantages of each test administration model are summarized in Table 1-1. Since this dissertation is interested in both CAT and ca-MST, these two testing models are elaborated on in Chapter 2.

Connection of Content Control with Test Fairness and Validity

Validity was once defined as a feature of a test device or instrument, but more recently the views and perceptions towards validity have changed (Kane, 2010). Messick's (1989) modern view on validity defines it as "integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment" (Messick, 1989, p.13). His well-accepted definition implies that validity is a broad issue, and in order to assess it, multiple source of evidence are required. This has been echoed in several chapters in the *Standards for Educational and Psychological Testing* (American Educational Research

Association [AERA], American Psychological Association [APA], & National Council Measurement in Education [NCME], 1999). These sources include evidential and consequential basis evidences. Messick (1989) states that one must use these evidences for two purposes: a) interpretation of test scores (i.e., describing someone's test performance) and b) use of the test (i.e., making decisions based on the test performance). Although some still prefer the traditional taxonomy (e.g., construct validity, content validity, internal validity, external validity) (see Sireci, 2007), Messick argues validity is a unified concept, and that there is only one type of validity, which is construct validity, and all other types of validity are branched under it (Messick, 1989).

Construct validity refers to the degree to which a test instrument measures what it is supposed to measure (Cronbach, 1971). The essential goal of assessing construct validity is to cover all pieces of the construct of interest (e.g., critical thinking) as much as possible, and to show that there is no construct underrepresentation or overrepresentation. Construct overrepresentation occurs when a test's content measures more than what it intends to measure and that is irrelevant with the targeted construct of interest. Construct underrepresentation occurs when a test's content misses some important pieces of the targeted construct of interest. This is where content control and construct validity interact. For example, a test that measures high school students' general math ability must include test items from all related content areas as required by the use of and interpretation of the scores, such as algebra, fractions, arithmetic, story based problems, functions, calculus, trigonometry, geometry, and possibly more. If the test content lacks items from one or more of the required areas, the insufficient content coverage results in construct underrepresentation, and thereby negatively impacts construct validity.

Consequently, this impacts the accuracy of measurement, interpretation of test scores, and thereby test fairness.

As stated before, linear tests have a unique advantage of high test designer control. Thus, it is easier to ensure that all examinees receive a sufficient number of test items from all content areas. In fact, in terms of this feature, they are even incomparable with other test administration models. In contrast, there is a disadvantage in adaptive tests because in adaptive tests examinees receive overlapping or non-overlapping test forms. Without proper algorithms for content balancing, the test might be terminated for a test taker before receiving items from all content areas. Thus, it requires more consideration and effort to ensure that each examinee receives equal and sufficient number of items from all content areas. To be able to fully control test fairness and measurement accuracy, and to draw valid score interpretations, more consideration should be given to the issue of content balancing requirements in adaptive testing administrations.

Since adaptive tests have become popular, many statistical procedures (e.g., item selection rule, stopping rule, routing method) have been proposed and tested under a variety of conditions. Many comparison studies have been conducted to investigate efficiency of these statistical procedures across CAT and ca-MST. Because it is directly related to validity, score interpretation and test fairness, non-statistical issues of adaptive tests, such as content balancing, have not been given enough attention. If content balancing is not met across the test forms received by examinees, other statistical conclusions made by test outcomes are not be comparable (Wainer et al., 2000). The goal of this study is to compare CAT and ca-MST in terms of content balancing control. This study aims to explore and compare the accuracy of outcomes produced by these two adaptive test approaches when strict content balancing is required.

Table 1-1. Main advantages and disadvantages of test administration models

Test Model	Advantages	Disadvantages
Linear Tests	Easy of test administration Low cost and effort for form construction No item pool requirement Item review High test developer control	Long test length Low measurement accuracy Late scoring and reporting Low test security No item review or skip
Computerized Adaptive Testing	Short test length Higher score accuracy High test security Quick scoring and reporting	High cost for item pool development Strong model assumptions Requires complicated mathematical design Not very easy of test administration
Computerized Multistage Testing	Intermediate test length Higher score accuracy (but less than CAT) High test security Quick scoring and reporting Allows item review and skip	High cost for item pool development Strong model assumptions Requires complicated mathematical design (but less than CAT) Not very easy of test administration

CHAPTER 2
LITERATURE REVIEW

Item Response Theory

Two widely used measurement theories for describing the relationships between observed variables (e.g., test items) and unobserved variables or latent traits (e.g., intelligence) are classical test theory (CTT) and item response theory (IRT). During the last century, CTT has been extensively used in the analysis of educational and psychological measurement. However, it is criticized by psychometric theoreticians and practitioners for producing test and population dependent outcomes and for focusing on true scores (Lord, 1980).

Another test theory approach of IRT is a strong mathematical theory, and describes the relationship between latent trait levels and responses to any given item. This relationship is described by item characteristic curves (ICC) (Crocker & Algina, 1986). An ICC (i.e., item response function) is an s-shaped logistic curve that indicates the expected probability of a particular response to an item conditional on the latent trait levels. Each item in the test task has a separate ICC. Figure 2-1 visually shows an example of an ICC for a binary item in which a score of 1 is a correct score. In this figure, the x-axis represents students' scaled versions of latent trait scores (θ), and the y-axis represents the probability of getting this item correct ($P(\theta)$) associated with each latent score. Latent trait score or level represents the ability of a student, and theoretically ranges from $-\infty$ to $+\infty$ but typically in practice ranges from -3 to 3 (Baker, 1992). As can be inferred from Figure 2-1, the probability of getting an item correct increases as the level of the latent trait increases.

There are three main properties for an item that shape the form of an ICC of a binary item. These item properties are item discrimination denoted by a, item difficulty denoted by b, and pseudo-guessing denoted by c. The item discrimination (i.e., slope or steepness parameter) is

an item's ability to differentiate examinees on the trait level (Embertson & Reise, 2000). The discrimination parameters typically ranges from 0.5 to 2 (Baker, 1992). A low *a* parameter makes the curve flatter and indicates that the probabilities are nearly equivalent across the different latent scores (Baker, 1992). A high *a* parameter makes the curve steeper. The item difficulty (aka threshold or location parameter) is the corresponding value on the latent trait scale when one has a 50% probability of getting the item correct. The difficulty parameter typically ranges from -3 to 3 (Baker, 1992). A low value of *b* indicates that the item is easy, and a high value of *b* indicates the item is difficult. As the *b* parameter for an item increases the location of its ICC shifts from left to right on the latent trait scale or vice versa. The last parameter is the pseudo-guessing parameter (i.e., lower asymptote), and it corresponds to the probability of getting the item correct for examinees that have infinitely low latent score.

Assumptions of Item Response Theory

IRT is considered a strong theory because it makes strong assumptions about the underlying data. There are three fundamental assumptions held by all IRT models: unidimensionality, local independence and monotonicity. Unidimensionality requires the items in the test to measure one and only one construct (e.g., math proficiency only). When the test measures more than one construct (e.g., both math and verbal ability), the assumption of unidimensionality is violated. Local independence states that there is no association between responses to any pairs of items in the test after controlling the effect of latent trait (Baker, 1992). This means that responses to the test items must be accounted for by the underlying latent trait only. One possible practical scenario that could lead to a violation of local independence is that one item provides a clue for another item. In such a case, responses to one item depend on the other item beyond the effect of the latent trait. The last assumption is monotonicity, which

assumes that the probability of correct response monotonically increases as the latent trait score increases (Embertson & Reise, 2000).

Dichotomous Item Response Theory Models

There are different types of IRT models such as dichotomous IRT models (i.e., for items having two score categories as true/false), polytomous IRT models (i.e., for items having more than two score categories as never/sometimes/always), and multidimensional IRT models (for items measuring more than one construct). The focus in this study is on dichotomous IRT models. There are three different latent trait models commonly used for dichotomous data. These are the one parameter model (i.e., Rasch model) (Rasch, 1960), two parameter model (Lord & Novick, 1968) and three parameter model (Birnbaum, 1968). The one parameter model (1PL) is the simplest IRT model and allows the difficulty parameter only to vary across items. The 1PL model defines the conditional probability of a correct response on item *i* for person *p* (X_{ip}=1) as

$$P\left(X_{ip} = 1 \middle| \theta_p\right) = \frac{e^{(\theta_p - b_i)}}{1 + e^{(\theta_p - b_i)}}, \tag{2-1}$$

where b_i is the difficulty parameter for item *i*, θ_p is the latent trait score for person *p*. It is important to note that under the 1PL model, the discrimination parameter is one and constant across all items. Due to this nature, the 1PL model makes an additional assumption of objective measurement. The two parameter (2PL) model has both item difficulty and item discrimination. The 2PL model defines the conditional probability of a correct response on item *i* for person *p* (X_{ip}=1) as

$$P\left(X_{ip} = 1 \middle| \theta_p\right) = \frac{e^{a_i(\theta_p - b_i)}}{1 + e^{a_i(\theta_p - b_i)}}, \tag{2-2}$$

where b_i is the difficulty parameter, a_i is the item discrimination for item i, and θ_p is the latent trait score for person p. As noted, the 2PL model allows each item to have separate discrimination. The three parameter (3PL) model has item difficulty, item discrimination and pseudo-guessing parameters that can vary across items. The 3PL model defines the conditional probability of a correct response on item i for person p (X_{ip}=1) as

$$P\left(X_{ip} = 1 \middle| \theta_p\right) = c_i + (1 - c_i)\frac{e^{a_i(\theta_p - b_i)}}{1 + e^{a_i(\theta_p - b_i)}}, \tag{2-3}$$

where b_i is the difficulty parameter, a_i is the item discrimination, c_i is the pseudo-guessing parameter for item i, and θ_p is the latent trait score for person p. The 3PL model allows each item to have both separate discrimination and pseudo-guessing parameter. In reality, each model is a subset of another model. For example, the 2PL model is the special form of 3PL model where the pseudo-guessing model is fixed at zero. The 1PL model is the special form of 2PL model where the discrimination parameter is constrained to one, and do not vary from item to item.

Item and Test Information in IRT

The dictionary definition of information is to learn about something or someone. The meaning of information in an IRT context is similar but has more statistical meaning (Crocker & Algina, 1986). The item information function of item i for a trait level under the 3PL model is defined as (Lord, 1980);

$$I_i(\theta) = a^2 \left[\frac{Q_i(\theta)}{P_i(\theta)}\right]\left[\frac{P_i(\theta) - c}{1 - c}\right]^2 \tag{2-4}$$

where $I_i(\theta)$ is the information of item i at trait level of θ, a_i is the item discrimination, $P_i(\theta)$ is the probability of correct response, $Q_i(\theta)$ is the probability of incorrect response ($P_i(\theta)$=1- $Q_i(\theta)$) for item i. As can be understood from the equation that in order to maximize the information, either

discrimination parameter or conditional probability of scoring a 1 should be increased. In case of the 2PL model, where c=0, Equation 2-4 turns into

$$I_i(\theta) = a^2\, P_i(\theta)\, Q_i(\theta). \tag{2-5}$$

In the case of the 1 PL, where a_i=1 and c_i=0, Equation 2-5 turns into

$$I_i(\theta) = a\, P_i(\theta)\, Q_i(\theta). \tag{2-6}$$

The test information for a trait level is simply obtained by summing all conditional item information and is defined as

$$TI(\theta) = \sum_{i=1}^{N} I_i(\theta), \tag{2-7}$$

where $TI(\theta)$ is the total test information for a N-item test. Figure 2-2 shows total test information functions for three different tests. Based on the figure, test one provides more information for people at latent trait around -1, test two provides more information for people at latent trait around 0, test three provides more information for people at latent trait around 1. In other words, test one is an easy test, test two is a medium difficulty test, and test three is a difficult test.

Item and test information (Birnbaum, 1968) play extremely important roles in IRT because they provide critical information about the extent to what performance an item and a test show (Embertson & Reise, 2000). The purpose of any test is to measure test takers' ability level more accurately (Hambleton & Jones, 1993). In order to achieve this, the standard error of measurement (SEM) at each latent trait score must be minimized. This is possible when the difficulty of items in the task match with latent traits of students because the SEM is the lowest around the item difficulty level (Embertson & Reise, 2000). Otherwise, the SEM will increase and items will not provide useful information about examinees' level latent traits. One of the key concepts in IRT is that the SEM is not constant across the different latent trait scores, therefore,

item information varies across the latent score scale. The amount of information an item provides peaks at its difficulty level, but is mainly determined by an item's discrimination (Baker, 1992). This means that as the discrimination increases the information an item provides increases. However, the location of the highest information point on the latent trait scale changes based on the difficulty level of the item. For example, items with high discrimination and low difficulty provide more information for low proficiency students, and items with both high discrimination and difficulty provide more information for high proficiency students. Adaptive testing makes use of information and SEM to target items onto examinees that maximize the former and minimize the latter.

Overview of Computerized Adaptive Testing

The first attempt of computerized adaptive testing was made with intelligence tests by Alfred Binet in 1905 (Wainer et al, 2000). Then, it was used during World War I for army recruitment purposes (DuBois, 1970). Even though these initial attempts were amateur as compared to today, they were invaluable in terms of burgeoning the field of alternative test administration. The late 1960s and early 1970s were the most critical years for CAT advancements. This is because Fredrick Lord, David Weiss, and their colleagues laid the basis of the first IRT based CAT applications. The interest in CAT has been especially intensified recently, and more than two hundreds journal articles have been published in the last twenty five years in three top educational measurement journals (i.e., *Journal of Educational Measurement, Applied Psychological Measurement,* and *Educational and Psychological Measurement*). The following parts in this chapter briefly describe principles and components of CAT.

An alternative name for computerized adaptive testing (CAT) is tailoring testing (Weiss & Kingsbury, 1984). This is because CAT selects each item based on test taker's current trait level estimate, thereby each examinee works at her own personalized test (van der Linden &

Glas, 2000). In general, the working principle of CAT is as follows. First, computer algorithms randomly administer an item (i.e., typically an item of medium difficulty) to an examinee. After her response to the first item, the computer estimates her latent score and selects the next item from the item pool that best matches with her current trait level. This means that as she gets the item correct, the computer gives a harder item, otherwise an easier item. This process continues until the stopping rule is satisfied. A flowchart in Figure 2-3 visually summarizes this.

Like the other test administration models, CAT has some great advantages. Perhaps, the most attractive advantage of CAT is high measurement precision compared to linear tests (van der Linden & Glas, 2000). The second advantage is lower test length. In fact, it has been shown that a well-designed CAT can reduce test length by 50% without any loss of measurement accuracy (Embertson & Reise, 2000). This is because each examinee receives items based on her ability level and does not waste her time with less or more difficult items. The third advantage of CAT is quick scoring and reporting. Examinees immediately learn their test score at the end of the test. Also, flexible test scheduling can be counted as another advantage associated with CAT.

There are some disadvantages associated with CAT. The first drawback is that CAT requires a large item bank. Item-writing is not easy work, and requires professionalism and money. A qualified item has to pass some certain criteria such as good discrimination and content appropriateness. Thus, this has to be done by content professionals. Furthermore, the estimated cost for a single qualified item ranges between $1,500 and $2,500 (Rudner, 2009). For instance, if we want to have a bank size of 500, it will cost between $750,000 and $1,250,000. In fact, this amount will increase as the bank needs maintenance. Compared to linear tests, this is a huge amount of money. The second drawback is that CAT is not very easy to implement (Yan et al., 2014). This is because it requires complicated software and fast computers to be available in

test centers. The third drawback is that CAT does not allow examinees to review the items that are already administered, and to skip any item during the test. Test takers have to respond all items and cannot go back to the previous item(s).

There are some chief components of CAT. These include item banks, item selection rules, ability estimation methods, item exposure controls, content balancing procedures, and stopping rules. The item bank consists of all pre-tested candidate items from all content areas, and psychometric properties of items (e.g., discrimination, difficulty, pseudo-guessing) in the bank are known in advance. Ability estimation method is the mathematical method used to calculate examinees' current and final latent trait scores. Item selection procedure consists of a computer algorithm used to select the next most useful item according to an examinee's current trait estimate. Item exposure control is a mechanism used to prevent items from being over-used. Content balancing rules are procedures used to ensure each examinee receives items from all content areas. Stopping rules are methods used to terminate the exam. Each component is detailed in the following section.

Item Bank

Item bank is a must in CAT administration. It is a collection of items that meet qualitative and quantitative requirements (Veldkamp & van der Linden, 2002). Both requirements are extremely important for an item bank design because they determine the quality of the bank, thereby affect CAT outcomes (Flaugher, 2000). Qualitative requirements represent content area specifications and item formatting. For instance, for a test to measure general reading ability, the item bank has to include many items that represent the following attributes: critical reading, understanding, vocabulary, paragraph comprehension, grammar. Furthermore, items should be carefully constructed and word count must be taken into account especially if the test is a speeded test. This is because students who receive too many items with a long word count will

have a disadvantage even if the items are easy (Veldkamp & van der Linden, 2002). Quantitative requirements represent psychometrics properties of the items in the bank. The items in the bank should be informative across different ability levels on the latent trait scale. Since the key feature of CAT is to adapt items to test takers' ability levels, there must be many items at different difficulty levels (e.g., easy, medium, difficult items). This will make the total item bank information function (IBIF) wider. It is important to note that the IBIF is determined according to the purpose of the test. For example, if the purpose of the test is to measure everybody equally (e.g., a norm-referenced test), then the shape of the IBIF should be rectangular distribution (Keng, 2008; Reckase, 1981). If the purpose of the test is to measure whether a test taker has sufficient knowledge (e.g., criterion-referenced test), then the shape of the BIF should be s-shaped distribution and peaked around the cut off score (Keng, 2008; Parshall, Spray, Kalohn, & Davey, 2002).

Item Selection Rule

Item selection rule is also a must in a CAT administration. The mathematical algorithm and/or working principle affects the performance of the selection rule and this impacts the outcomes. Most commonly used item selection rules are summarized in this section. The first proposed item selection method was Owen's item selection rule (Owen, 1975). This procedure is based on Bayesian approach and selects the next item that reduces the variance of the posterior distribution and keeps selecting items until the variance of the posterior distribution reaches the specified level. One can refer to Owen (1975) for more technical details. An alternative procedure is maximum information proposed by Brown and Weiss (1977). Since the item properties in the bank are known in advance, the amount of information each item provides at the gridded latent abilities on the trait scale is also known in advance. This method of item selection selects the next item that maximizes the information at the current provisional trait. This method

has been very popular and extensively used in the CAT literature. However, it has a limitation that items with high item discrimination will provide more information and be more likely over-selected, especially at early stages of CAT (van der Linden & Glas, 2000). That jeopardizes test security. Thus, this method should be used with a caution (i.e., item exposure rate should be controlled if it is of interest). One can refer to Brown and Weiss (1977) or Lord (1980) for the technical details.

Another item selection method is *b*-stratified method proposed by Weiss (1973). This method stratifies item bank into group of items based on difficulty levels, and selects the next item that provides the highest information from the stratifications. However, a limitation of this method is that information is mainly determined by item difficulty not item discrimination. One can refer to Weiss (1973) for technical details. Another item selection method is Kullback-Leibler information (Chang & Ying, 1996). In this method of item selection, the Kullback-Leibler (KL) information, which is the distance between two ability points on the latent trait score, is first estimated across the number of k items already administered for an examinee. Then, the next item that minimizes the KL information (i.e., best matching item with the current latent score) is selected. One can refer to Chang and Ying (1996) or Eggen (1999) for technical details. Another item selection rule is *a*-stratified method proposed by Chang and Ying (1999). This method overcomes problems with maximum information and *b*-stratified method by stratifying the item bank by discrimination parameters. As stated, items have high discrimination are consistently selected with the maximum information method. This method selects items from different item groups and solves item exposure problem in the maximum information method. One can refer to Chang and Ying (1999) for technical details.

Another item selection method is weighted information criteria proposed by Veerkamp and Berger (1994). This method is similar to the maximum information method but specifies weights for item information across the items and groups items based on the weighted information, and selects the next item from these grouped items. One can refer to Veerkamp and Berger (1994) for technical details. Furthermore, there are many new rules has been proposed in last two decades such as proportional method (Barrada, Olea, Ponsoda, & Abad, 2008), and revised rules Kullback-Leibler function weighted by likelihood (Huang & Ying, 1996). One can refer to the cited articles for more description.

Ability Estimation

Like the item selection rule, ability estimation method is also a must in a CAT administration. There are two main ability estimation method groups: Bayesian methods and non-Bayesian methods. The most commonly used non-Bayesian method is maximum likelihood estimation (MLE) (Hambleton & Swaminathan, 1985). MLE begins with some a priori value (e.g., starting theta value) for the ability of the examinee and calculates the likelihood function with given response pattern (e.g., 0101100111) for an ability level. The likelihood function estimates the probability of having that response vector and finds the value of theta that most likely results in that observed pattern. For example, if a person got many items wrong (e.g., response vector of 000010001), the likelihood function will tell us that this response pattern belongs to someone that has very low latent trait. Then, MLE finds the point that maximizes the likelihood of an examinee's item response pattern, and goes to the corresponding score on the latent trait scale. This score is the estimate of latent trait for a test taker. First advantage of this method is that MLE is mathematically very easy procedure to use. Another advantage is that since item parameters are known in advance in CAT, estimations with MLE are unbiased compared to linear tests in which parameters of items are unknown (Wang & Vispoel, 1998).

One big disadvantage associated with MLE is that it does not provide an estimate for examinees that get all items right (i.e., perfect scores) or wrong (i.e., zero scores). This causes the likelihood function to infinitely increase and it becomes impossible to find the highest value on the likelihood function. This is a serious problem especially at early stages of CAT (Keller, 2000), and it is not recommended for short test length CATs (Wang & Vispoel, 1998). One can refer to Hambleton and Swaminathan (1985) and/or Lord (1980) for technical details.

Three Bayesian methods commonly used in CAT are maximum a posteriori (Lord, 1986), expected a posteriori (Hambleton & Swaminathan, 1985), and Owen's method. Maximum a posteriori (MAP) is similar to MLE but MAP specifies a prior distribution and then multiples this prior distribution by the likelihood function and then, does the same thing with the MLE. Most often, the prior distribution is chosen from a normal distribution. One big advantage of MAP is that it provides estimates for perfect and zero scores and outperforms MLE (Wang & Vispoel, 1998). Another advantage associated with MAP is that ability estimation can be done even after the first item is administered in CAT and does not assume normality (Keller, 2000). One can refer to Hambleton and Swaminathan (1985) and/or Keller (2000) for technical details.

Another Bayesian method is expected a posteriori (EAP). EAP also specifies a prior distribution but unlike the highest point on the likelihood function found in MAP, it finds the mean of the posterior distribution which represents the estimate of the latent trait. Unlike the MLE and MAP, it is a non-iterative procedure, so easy to implement. Also, it does not assume normal prior distribution as MAP, and EAP outperforms MLE and MAP meaning that produces lower standard error and bias than other CAT ability estimation methods. One possible disadvantage with EAP is that if inappropriate prior distribution is specified, this affects the accuracy of outcomes. Also, for more accurate results, EAP requires at least 20 items in CAT,

otherwise produce biased results than others (Wainer & Thissen, 1987). One can refer to Wainer and Thissen (1987) and/or Hambleton and Swaminathan (1985) for technical details. The last commonly used Bayesian method is Owen's method (Owen, 1975). Like the other Bayesian methods, this method also specifies a prior estimation and then does the same thing with MLE but the difference is that the prior distribution is updated after each administrated item. It does this by specifying the prior distribution so that the mean and standard error of the prior distribution is equal to the mean and standard error of the posterior distribution in the most recent step (Keller, 2000). As the other Bayesian methods, it also allows ability estimates for examinees with perfect or zero scores. This method is also a non-iterative procedure so easy to implement. However, it produces high bias for ability estimates and is the least recommended Bayesian method in a CAT environment (Wang & Vispoel, 1998). One can refer to Owen (1975) for technical details.

Item Exposure Control

By nature, all item banks have some good and poor items for a particular trait score (Revuelta & Ponsoda, 1998). As stated, the main goal of CAT is to find best item for a current latent trait estimate. Good items (i.e., items provide more information at a particular latent trait score) have higher likelihood of being selected, whereas, bad or poor items has lower chance of being selected. Thus, good items are selected again and again for different examinees, and poor items remain as unused. In such cases, some examinees might memorize these items, and share them with other test takers. This jeopardizes test security (Yi, Zhang, & Chang, 2006). In order to increase test security, item exposure rate should be controlled. There are more than thirty item exposure controls proposed in the last four decades. One can refer to Georgiadou, Triantafillou and Economides (2007) for full list and more details. In this study, we will summarize the most popular and commonly used procedures. They are a) 5-4-3-2-1 strategy (McBride & Martin,

1983), b) randomesque strategy (Kingsbury & Zara, 1989), c) Sympson-Hetter strategy (Sympson & Hetter, 1985), and d) *a*-Stratified strategy (Chang & Ying, 1999).

As stated, item exposure is a serious problem especially at early stages of CAT because at the beginning item selection methods might tend to select best informative items. In the 5-4-3-2-1 method, after the starting item and initial ability estimate, instead of selecting the best informative item for a current ability estimate, the next first item (i.e., examinee's second item in the test) is randomly selected from the top five informative items. The next second item is randomly selected from the four best informative items, the next third item is randomly selected from the four best informative items, the next fourth item is randomly selected from the two best informative items, and then system returns to normal flow and the next fifth item and subsequent items are selected to be best informative at the current ability estimate. Randomesque strategy is very similar to the 5-4-3-2-1 strategy in terms of attempting to minimize item overlap among the examinees. The difference is that instead of decreasing number of informative item groups, randomesque strategy constrains item selection to choose items always from the same number of best informative item groups (e.g., 2, 5, 10). For example, after starting item and initial trait estimate, the next first item is selected from top ten informative items, the second next item is also selected from the top ten informative items, and so on. These two methods are useful at early stages of CAT but do not control exposure rates over all items in the bank. Another strategy is Sympson-Hetter (SH) which controls the exposure rate of all items in the bank, thereby, has been the most popular method so far. The SH is a probabilistic method, requires to set up a pre-determined exposure rate value for the items (e.g., r_i). This method calculates number of times an item administered as a proportion (e.g., ranges from 0 to 1) and record it. A proportion value of 1 means that the item is selected for everybody, a proportion value of zero means that the item is

never selected. For example, the desired exposure rate for the items in the CAT version of Graduate Record Examination (e.g., GRE) was 0.20 (Eignor, Stocking, Way, & Steffen, 1993). This means that maximum 20% of the examinees were receiving an item. If the proportion for item *i* is lower than r_i, the selected item is also selected for another examinee, if it is equal to or higher than the r_i, returns to the bank. The last method is *a*-Stratified strategy (*a*-STR). This methods ranks items from low discrimination to high, and stratifies items into the bank based on discrimination parameters (e.g., *K* strata). For an ability estimate, one of the items into a strata is selected. At the early stage of CAT, since little information is known for ability estimates, items are selected from the stratas that have low discriminating items, later on since more information is known, items are selected from the stratas that have high discriminating items.

Content Control

As stated, for a successful and valid CAT, it is a must to control content balancing so that each test taker takes equal and/or similar number or percentage of items from each content area. In this study, we will briefly summarize the four most commonly used content control methods. First one is the constrained computerized adaptive testing (Kingsbury & Zara, 1989) which has been used in real CAT applications (CAT-ASVAB, and Computerized Placement Tests) over the years (Chen & Ankenmann, 2004; Ward, 1988). As it is elaborated, the main idea in item selection is that the selected item is the one that maximizes the information at the current latent trait estimate. The constrained computerized adaptive testing (CCAT) tabulates item informations according to pre-specified content areas, and seeks the content area in which the number/percentage of selected items is farthest below the targeted percentages (Kingsbury & Zara, 1989). Then, the most informative item is selected from this content area. The CCAT method goals to reach targeted level of test specifications. This method is very simple and efficient to meet the desired content area distributions as long as the item pool is large enough.

However, it has a limitation that sequence of content areas is predictable especially at early stages of CAT. As stated in Chen and Ankenmann (2004), after this method is used for a while later, examinees might predict the next content area and this content order knowledge might affect the probability of getting the item correct (Chen & Ankenmann, 2004). Another popular method is modified CCAT (MCCAT) proposed by Leung, Chang and Hau (2000) to overcome deficiencies of CCAT. This method basically does the same thing but, items are selected from any content area instead of from the one that has the lowest exposure rate and deal with potential content order knowledge. Even though this effort, both CCAT and MCCAT do not provide efficient content control when test length varies across examinees (Chen & Ankenmann, 2004). Another content control of weighted deviation modelling (WDM is a mathematical strategy and, proposed by Stocking and Swanson (1993). The WDM specifies weights for content areas and as the CAT continues, for each content area calculates deviation from the targeted test specification, and then forces item selection to select items from mostly deviated content area (e.g., the one minimizes weighted distance). The method of maximum priority index (MPI) is proposed by Cheng and Chang (2009), and very popular method recently. In this method, instead of item information itself, the information an item provides for a theta level is multiplied by a weight. The number of weights are equal to the number of content areas, and items in the same content area have equal weight. Furthermore, items in major contents have higher weights, items in minor contents have lower weights. The final equation (e.g., product of weight by information) is called the priority index. Then, the item that maximizes the priority index is selected as the next item (Cheng & Chang, 2009). For more technical detail, the readers can refer to the original studies.

Stopping Rule

A stopping rule determines the length of CAT, and is another essential component in CAT (Babcock & Weiss, 2009). Unlike the other components, stopping rule is less complicated. There are two different commonly used stopping rules in CAT as a) fixed test length and b) varying test length. Fixed test length requires all examinees to receive the same number of pre-specified items. For example, if the specified test length is 60, the test is terminated for a test taker after she responded her 60th item, regardless of the level of measurement accuracy. Another stopping rule is varying test length (aka minimum information criteria or minimum standard error-SE). This rule terminates the test for a test taker when the target level of precision is obtained (Segall, 1996). For example, if the pre-specified value of SE is 0.3, the test stops for a test taker when her latent trait is calculated with a standard error of 0.3. It has been an argument over the years that varying test lengths produce biased results than fixed tests (see Babcock & Weiss, 2009; Chang & Ansley, 2003; Yi, Wang, & Ban, 2001). Babcock and Weiss (2009) tested both rules under a variety of conditions and concluded that varying tests produce equal or slightly better results than fixed tests (Babcock, & Weiss, 2009). Their huge simulation study showed that a) CATs lower than 15 or 20 items do not produces good estimates of latent trait, b) CATs higher than 50 items do not provide meaningful gain in estimates of latent trait, c) recommended value of SE is 0.315 for varying CATs (Babcock & Weiss, 2009).

Furthermore, it is also possible to combine these methods. For example, one can design a CAT that requires a target measurement precision by constraining minimum and maximum number of items (see Thissen & Mislevy, 2000). This can be efficient to prevent both very early termination and test fatigue. It is important to note that like item bank design, stopping rule also depends on the purpose of the tests (Segall, 2005). For example, if the target is to gain equal measurement precision across all examinees for interpretation and/or decision purposes (e.g.,

personality test), then varying test should be used. If the test a placement test for hiring decisions and test fatigue is a concern, fixed test length should be used.

Overview of Computerized Multistage Testing

The multistage testing has been proposed as an alternative test administration method to CAT. Over the years, it has been called by different names such as two-stage testing (Kim & Plake, 1993), computerized mastery testing (Lewis & Sheehan, 1990), computer adaptive sequential testing (Luecht, 2000), and bundled multistage testing (Luecht, 2003). Although operational use of multistage testing is relatively new compared to linear tests and CAT, the MST idea is not new (Yan et al., 2014). Early MST designs date back to the 1950s (see Angoff & Huddleston, 1958). The first versions of MSTs were on paper and pencil format, and there was no adaptation from one point to another (see Cronbach & Glaser, 1965; Lord, 1971 & 1974; Weiss, 1973). However, since the research on CAT eclipsed MST, adaptive version of multistage testing, which is the interest in this study, has been proposed for use (Keng, 2008; Mead, 2006). Most recently, the adaptive version of multistage testing is called computer adaptive multistage testing or computerized multistage testing (ca-MST), and has gained its popularity recently, and more than one hundred journal articles have been published in the last twenty years. Also, the number of operational examples has increased recently (e.g., The Massachusetts Adult Proficiency Test, GRE, The Law School Admission Council, The Certified Public Accountants (CPA) Examination). In terms of test feature, it can be said that ca-MST is a combination of linear tests and CAT.

The ca-MST terminology includes some special terms not used in other testing procedures. These include module, stage, panel, routing, path and test assembly. The ca-MST is comprised of different panels (e.g., a group of test forms), panels are comprised of different stages (e.g., division of a test), and stages are comprised of pre-constructed item sets (i.e., in ca-

MST literature, each set is called a module) at different difficulty levels (Luecht & Sireci, 2011). This means that at each stage some of the modules are easier and some of them are harder. In other words, some modules provide more information for low proficiency test takers, some provide more information for high proficiency test takers. There is almost always one module in stage one which is called the routing module which is used to establish test taker proficiency level. The test taker moves to the next module on the test based on the performance on the routing module. The number of the stages and the number of the modules in each stage, and the number of the items in each module can vary from ca-MST to ca-MST.

Figure 2-4 shows an example of simplest possible ca-MST structure with two stages and two modules in stage two. This structure is called the 1-2 panel design. Two numbers in the structure name means there are two stages, each number represents total number of modules at that stage. As shown in the figure, there are two possible pathways that a test taker might draw (e.g., Routing-Easy and Routing-Hard). Figure 2-5 shows a more complex ca-MST design with 3 stages and 3 modules in stage 2 and 3. This structure is called the 1-3-3 panel design. Similarly, the three used numbers in the structure name means a total number of stages, and each number in the triplet represents the number of modules at that stage. There are seven possible paths in this type of ca-MST design. These are: Routing-Easy-Easy, Routing-Easy-Medium, Routing-Medium-Easy, Routing-Medium-Medium, Routing-Medium-Hard, Routing-Hard-Medium, Routing-Hard-Hard. As can be understood from the figure, the pathways from a module to another module in the next stage that is not adjacent to the current module are ignored. For example, if a student receives easy module at the stage 2, even if she shows very high performance on this module, the student is not permitted to receive the hard module at the stage 3. The strategy of disallowing extreme jumps among the module is highly common in many ca-

MST designs (Luecht, Brumfield, & Breithaupt, 2006), and prevents aberrant item responses and inconsistent response patterns (Davis & Dodd, 2003; Yan et al., 2014).

Both Figure 2-4 and 2-5 display an example of one panel only. Whereas, a computerized multistage panel is referred a collection of modules (Luecht & Nungester, 1998). In order to control panel exposure rate (and thereby module and item), ca-MST designs consist of multiple panels that are similar to each other, and each test taker is randomly assigned to one of them. Figure 2-6 shows an example of multiple panels. Panel construction procedure will be described later on.

In general, the working principle of ca-MST is as follows. After assigning a test taker to a panel, unlike individual items in CAT, ca-MST starts with a routing module (e.g., a set of 5 or 10 items). After the routing module, the first stage of ca-MST, the computer calculates the test taker's latent trait score. Then, based on her current latent score, the computer selects one of the pre-constructed modules in the second stage, and routes the test taker to the appropriate module. For example, if her performance on the routing module is a high, she receives hard module in the second stage, otherwise easier module is selected. After she completes the second stage, again, computer calculates her current latent trait score, and routes her to the most appropriate module in the third stage. This process continues until the test taker completes all stages. It is very clear that the main distinction between CAT and ca-MST is that there is an item level adaptation in CAT in contrast to the module level adaptation in ca-MST.

Like the other test administration models, the ca-MST also has some advantages and disadvantages over other test administration models. Perhaps, the most attractive advantage of ca-MST over CAT is that ca-MST is more flexible in terms of item review and item skipping. ca-MST allows examinees to go back to the previous items within each module, and to skip any

item as well. However, examinees are not allowed to go back to the previous stage(s), and review items in the previous module(s). Compared to linear tests, the ability to go back to the limited number of previous items is a drawback but compared to the CAT, it is a remarkable feature. There is a well-known motto that the first answer is always correct, and students should always trust their initial responses (van der Linden, Jeon, & Ferrara, 2011). However, a research by Bridgeman (2012), an ETS practitioner, proved that common belief is a superstition, and there was an increase in measurement of student abilities when examinees had chance to review the items (Bridgeman, 2012).

In terms of test length, ca-MST stays between linear tests and CAT. This means that ca-MST produces shorter test length than linear tests but longer test length than CAT (Hendrickson, 2007). In terms of test design control (e.g., content area, answer key balance, word counts etc...), ca-MST allows more flexibility than CAT but less than linear tests. In terms of measurement efficiency, ca-MST produce much better measurement accuracy than linear tests (Armstrong, Jones, Koppel, & Pashley, 2004; Patsula, 1999) but equal or slightly lower accuracy than CAT (Armstrong et al., 2004; Kim & Plake, 1993; Patsula, 1999).

However, some of the drawbacks to CAT such as high cost of item bank creation still remain as an important concern in ca-MST. Also, while CAT can stop at any point (if stopping rule is satisfied), and reduce the test time, due to the module level adaptation, the ca-MST continues until one completes all stages (Zheng, Nozawa, Gao, & Chang, 2012). By taking into account the advantages and disadvantages of ca-MST, it can be said that ca-MST is a highly promising test administration model because it “strikes a balance among adaptability, practicality, measurement accuracy, and control over test forms” (Zenisky, Hambleton, & Luetch, 2010, p.369). Like the CAT, ca-MST has some chief components as test design and

structure, routing method, ability estimation, content control and test assembly, exposure control. Since the same ability estimation methods (e.g., ML, MAP, and EAP) are used in ca-MST, and conclusions in CAT are the same in ca-MST, ability estimation section is not presented again. Remaining components are elaborated below.

Test Design and Structure

Prior to an ca-MST administration, one must determine panel design or structure. This includes determining number of panels, number of stages, number of modules in each stage, number of items in each module, and total number of test items. These steps are summarized in the following sections.

Number of panels

As stated before, ca-MST comprises parallel test forms called panel, and assignment of test takers to the panels occurs randomly (Luecht, 2003). Having multiple panels helps to reduce panel, module and item exposure rate, and prevent items being overused. This is very critical to increase test security. This issue is elaborated under “exposure control”. Depending on the importance of the exam (e.g., high stake or low stake), the number of panels changes but in both operational examples and simulation studies it usually varies from one to forty (Yan et al., 2014). Panel construction procedure is handled under “Content Control and Automated Test Assembly”.

Number stages and modules

As stated before, modules in each stage differ in average difficulty levels, but the number of items and proportion of content specifications are the same across the modules in a stage (i.e., not necessarily be the same in modules at different stages). There are several examples of ca-MST design configurations in the literature such as 1-2 (see Wang, Fluegge, & Luecht, 2012), 1-3 (see Schnipke & Reese 1999; Wang, Fluegge, & Luecht, 2012), 1-2-2 (see Zenisky, 2004), 1-3-3 (see Keng & Dodd, 2009; Luecht et al., 2006; Zenisky 2004), 1-2-3 (see

Armstrong & Roussos 2005; Zenisky 2004), 1-3-2 (see Zenisky,2004), 1-1-2-3 (Belov & Armstrong, 2008; Weissman, Belov, & Armstrong, 2007), 1-5-5-5-5 (Davey & Lee 2011), 1-1-2-3-3-4 (Armstrong et al. 2004), 5-5-5-5-5-5 (Crotts, Zenisky, & Sireci 2012). The two stage design is the simplest and most widely used in both operational applications (e.g., revised version of GRE) and simulation studies. This is because there is only one adaptation point in this configuration but this property also brings it to the disadvantage of higher likelihood of routing error (Yan et al., 2014). Some suggested using an additional module called "recovery module" when necessary, but it was not found very attractive (see Schnipke & Reese 1999). Thus, as the previous studies showed, although the test complexity increases, adding more stages and modules into each stage, and/or allowing more branching increases test outcomes (Luecht & Nungester 1998; Luecht, Nungester, & Hadadi 1996). This is more likely due to having more adaptation points. Armstrong, et al (2004) and Patsula and Hambleton's (1999) huge simulation studies displayed that having more than four stages does not produce meaningful gain in test outcomes, and two or three stages with two or three modules at each stage are sufficient for a ca-MST administration (Armstrong, et al. 2004; Patsula & Hambleton, 1999; Yan et al., 2014).

Number of items

The ca-MST is a fixed-length test so, it is necessary to decide the total test length for a test taker. The total test length varies from ca-MST to ca-MST (e.g., from 10 to 90). Then, it is necessary to decide the number of items in each module. It is possible to assign different number of items into the modules at different stages. However, based on the previous research a) adding more items into the tests increases reliability of test scores and thereby measurement accuracy (Crocker & Algina, 1986), b) varying number of items in modules at different stages do not affect test outcomes (Patsula & Hambleton, 1999), c) increasing number of items in the routing

module have positive effect on test outcomes (Kim & Plake, 1993), d) under the same number of total items, having shorter modules is better due to allowing more adaptation points.

Routing Method

Routing method is analogous to item selection method in CAT. Since there is a module level adaptation in ca-MST, routing method is used to decide the subsequent modules that an examinee will receive (Luetch, 2003). The efficiency of routing method affects the pathway a student draws during the test, so misrouting impacts the test outcomes, and thereby usefulness of ca-MST (Lord, 1980). Thus, it is tremendously important component in ca-MST. The routing methods are considered into two categories as dynamic rules and static rules (Yan et al., 2014). The dynamic method of maximum Fisher's information (MFI) (Lord 1980; Thissen & Mislevy 2000) is an information-based method, and analogous to the MI item selection method in CAT. That is that computer algorithm calculates examinee's ability level based on the previously administered module(s), and then selects the module that maximizes information at his/her current ability estimate (Weissman et al., 2007).

Both two static methods define cut points (e.g., routing points or upper and lower bounds) for latent traits when routing examinees to the modules but they differ in defining cut points. First static rule of module selection method is the approximate maximum information method (AMI) (Luecht, 2000) which is mainly used in criterion-referenced test administration (Zenisky, 2004). In AMI routing method, computer algorithm calculates cumulative test information function (TIF) based on the previously administered modules, and the TIFs of the current alternative modules (e.g., easy, medium, hard modules). Then, it adds each alternative module's TIF to the current cumulative TIF, separately, and defines the intersection points of TIFs as the cut points (e.g., if the two intersection points of three TIFs are -1 and 1, cut points are -1 and 1). Finally, computer routes examinees to the one of the alternative module that provides

the highest information for the provisional latent trait of an examinee. (e.g., if $\theta \leq -1$ examinee is routed to easy module, if $-1 \leq \theta \leq 1$ examinee is routed to medium module, if $\theta \geq 1$ examinee is routed to hard module).

Second static rule of module selection method is the defined population interval (DPI) or number-correct (NC), which is mainly used in norm-referenced test administration (Zenisky, 2004). The main goal in this method is to route specific proportion of people to the modules to ensure equal or highly close number (or proportion) of people draws each possible pathway (e.g., 33% of examinees routed to easy module, 34 % of examinees routed to medium module, 33% of examinees routed to hard module). For example, in a 1-3-3 panel structure, first the θ values corresponding to the 33rd and 67th percentiles of cumulative θ distribution are calculated (e.g., if the latent scores are normally distributed, θ-scores for the 33rd and 67th percentiles are -0.44 and 0.44, respectively). These are actually defined as the cut points in DPI method. Then, these cut points are transformed to corresponding estimated true scores. Finally, examinees are routed to the one of the alternative modules based on the number of correct responses they got on the current module (e.g., out of 10 items, people got 6 or less items correct are routed to easy module, people got 7 or 8 items correct are routed to medium module, people got 9 or 10 items correct are routed to hard module) (see Zenisky et al., 2010 for more details).

Compared to MFI and AMI method, the DPI is fairly straightforward to implement but it requires to an assumption on the distribution of theta scores to specify cut points prior to the test administration. Misrouting is more likely to occur when there is a huge discrepancy between the actual and assumed distributions. However, previous research showed that in terms of routing decisions, DPI or NC is very practical and sufficient for a ca-MST design (Weissman et al.,

2007). However, the choice of routing method is mainly determined according to the purpose and consequences of the test (Zenisky et al., 2010).

Automated Test Assembly, Content Control and Exposure Control

The ca-MST is designed in such a way that test takers receive pre-constructed modules based on their performance on the previous module (Armstrong & Roussos 2005). This means that each subsequent module has to match with the current ability of a taker. Thus, items must be carefully grouped into the modules. The most critical consideration here is to group items in the item bank into the modules based on the information functions. Some used trial-error method and manually assembled modules (see Davis & Dodd, 2003; Reese & Schnipke, 1999), but it is quite difficult to satisfy all constraints (e.g., number of items, controlling content area) and to create parallel panels with manually assembled tests. Thus, it is always best to use a better strategy because the automated test assembly (ATA) is a must in ca-MST (Luecht, 2000).

"The automated test assembly involves the use of mathematical optimization procedures to select items from an item bank for one or more test forms, subject to multiple constraints related to the content and other qualitative features" (Luetch, 2003, p.7). The optimum solution of ATA procedure ensures that items in modules and panels meet the desired constraints such as difficulty level, content control, word count, item and test overlap, item format etc… As stated before, one has to decide ca-MST panel structure, total test length for a test taker and total number of panels prior to the ca-MST administration. Next, based on the ca-MST structure, the test designer has to determine the number of items in each module, and to pull a group of items from the given item bank that meet all desired constraints. There are some software used to solve ATA problems in ca-MST studies such as IBM CPLEX (ILOG, Inc., 2006), CASTISEL (Luecht, 1998). After solving the ATA, the test designer has to place items into the modules and these modules into the panels. There are two approaches used to assign modules into the panels,

bottom-up and top-down (Luecht & Nungester, 1998). In bottom-up approach, all modules are built so as to meet module level specifications such as content, target difficulty. It is possible to think of each module as a mini test (Keng, 2008). Since modules constructed with this strategy are strictly parallel, modules are exchangeable across the panels. In bottom-down approach, the modules are built based on test level specifications. Since the modules are dependent and not parallel, not exchangeable across the panels.

Statement of the Problem

Large scale tests in educational and psychological measurement are constructed to measure student ability using items from different content areas related to the construct of interest. This is a requirement for a validity because the degree of validity is related to showing that test content aligns with the purpose of the use of the test and any decisions based on test scores (Luecht, de Champlain, & Nungester, 1998). This means that a lack of test content from one or more areas severely impacts validity. It is well known that even if score precision is very high as would be the case through proper adaptive algorithms in CAT or ca-MST, this does not necessarily ensure that the test has valid uses (Crocker & Algina, 1986). In other words, high score precision might not represent the intended construct if the content balancing in not ensured. Furthermore, if all students are not tested on all aspects of the construct of interest, test fairness is jeopardized. For these reasons, any test form adapted to an examinee has to cover all related and required sub-content areas for validity and fairness purposes.

As discussed in the literature review, both CAT and ca-MST have a long history of use. Much research has been conducted on CAT including the proposal of new item selection methods (e.g., Barrada et al., 2008; Chang & Ying, 1996), stopping rules (e.g., Choi, Grady, & Dodd, 2010), and exposure control methods (e.g., Leung et al., 2002; van der Linden & Chang, 2005). Researched areas of ca-MST include proposals for new routing methods (Luetch, 2000;

Thissen & Mislevy 2000), test assembly methods (Luetch, 2000; Luecht & Nungester, 1998), and stage and module specifications (Patsula, 1999). Furthermore, many comparison studies have been conducted to explore efficiency of CAT versus ca-MST in terms of different test outcomes (see Davis & Dodd, 2003; Hambleton & Xing, 2006; Luecht et al., 1996; Patsula, 1999). However, the main focus in the comparison studies was the statistical components of CAT and/or ca-MST, and little consideration has been given to the non-statistical aspects of adaptive tests such as content balancing control (Kingsbury & Zara, 1991; Thomasson, 1995). This is critical because selection of items in CAT and/or assigning items into the modules in ca-MST is confounded by content balancing concerns (Leung, Chang, & Hau, 2003). In order to reap the benefits of an adaptive tests over linear tests, content balancing has to be strictly met (Luecht et al., 1998), otherwise, test scores of students will be incomparable, and test validity will be questionable (Segall, 2005).

So far, many aspects of CAT and ca-MST have been compared under varying conditions including test length (Kim & Plake, 1993), item exposure rates (Davis & Dodd, 2003; Patsula, 1999), item pool characteristics (Jodoin, 2003) and pool utilization (Davis & Dodd, 2003). The common findings in these studies was that due to the item-level adaptation and/or more adaptation points, CAT produces better accuracy of ability estimation (Yan et al., 2014). However, it is consistently asserted in several studies that a major advantage of ca-MST is that it controls for content better than CAT (see Chuah, Drasgow, & Luecht 2006; Linn, Rock, & Cleary 1969; Mead 2006; Patsula & Hambleton 1999; Stark & Chernyshenko 2006; van der Linden & Glas 2010; Weiss & Betz, 1974; Yan et al., 2014). Yet, the literature does not contain a study that specifically compares CAT with ca-MST under varying levels of content constraints to verify this claim. It is obvious that due to the feature of test assembly, ca-MST can easily meet

content constraints. However, it is still unknown what we lose by administering ca-MST versus CAT when the two are different in how they select items from the same item bank to meet varying levels of content control goals. This dissertation aims to explore the precision of test outcomes across the CAT and ca-MST when the number of different content areas is varied across a variety of test lengths. The study seeks answers to the following research questions:

1. How will the test outcomes be impacted when number of content area and test length are varied within each testing model (e.g., CAT, 1-3 ca-MST, 1-3-3 ca-MST)?

2. How will the test outcomes be impacted when number of content areas is varied under different panel designs (1-3 vs 1-3-3) and different test lengths (24-item and 48-item test length) on the ca-MST?

3. How will the test outcomes be impacted on the CAT and ca-MST under the combination of the levels of test length and content area?

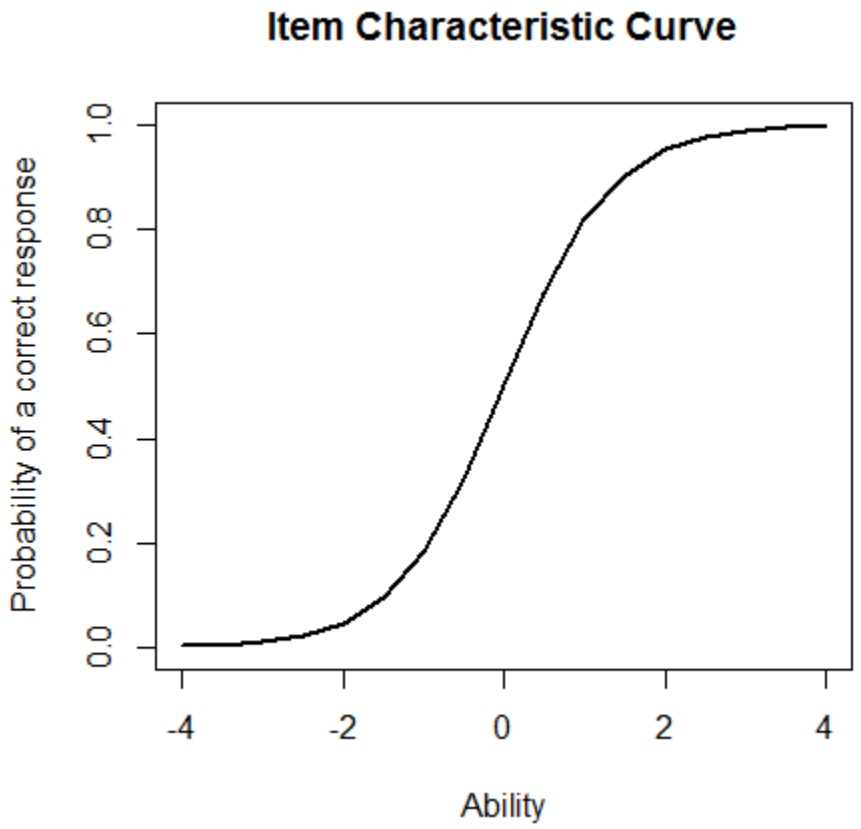

Figure 2-1. An example of item characteristic curve (ICC).

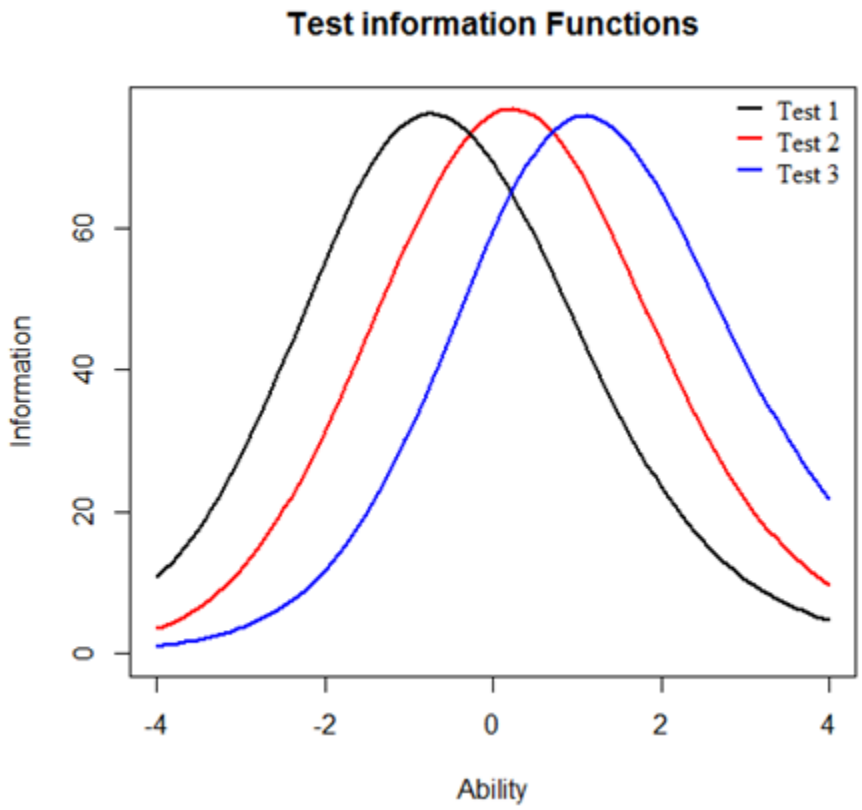

Figure 2-2. Total test information functions for three different tests.

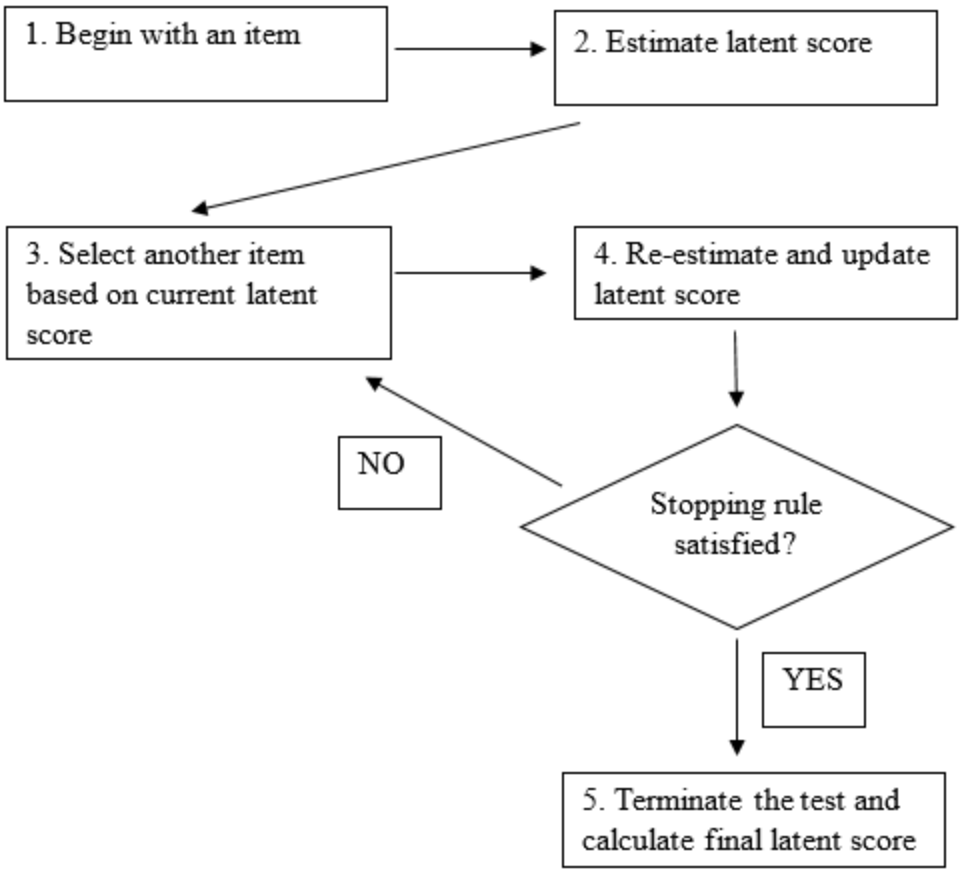

Figure 2-3. The flowchart of computerized adaptive testing.

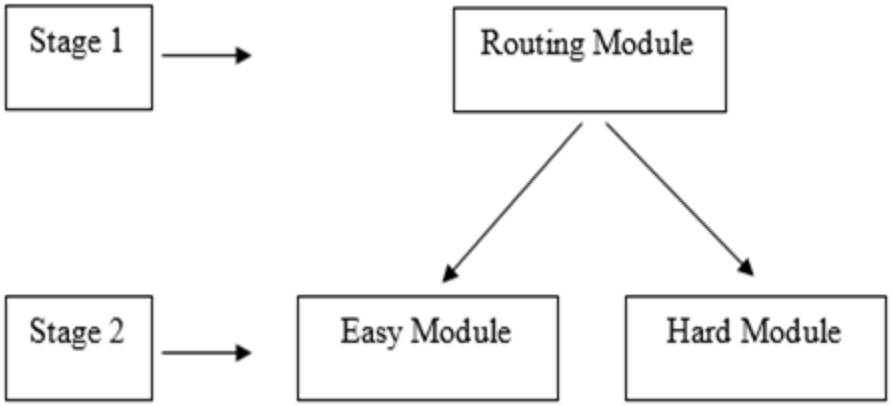

Figure 2-4. An example of 1-2 ca-MST panel design.

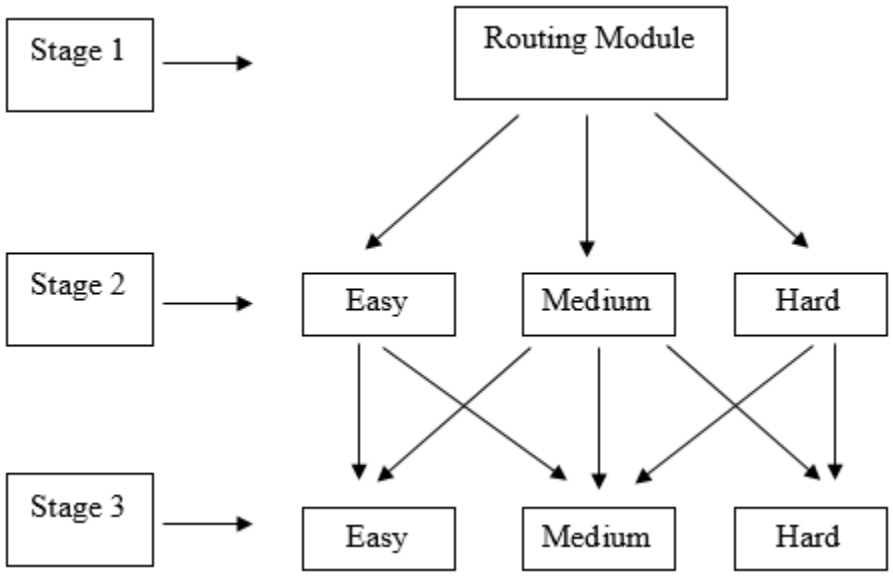

Figure 2-5. An example of 1-3-3 ca-MST panel design

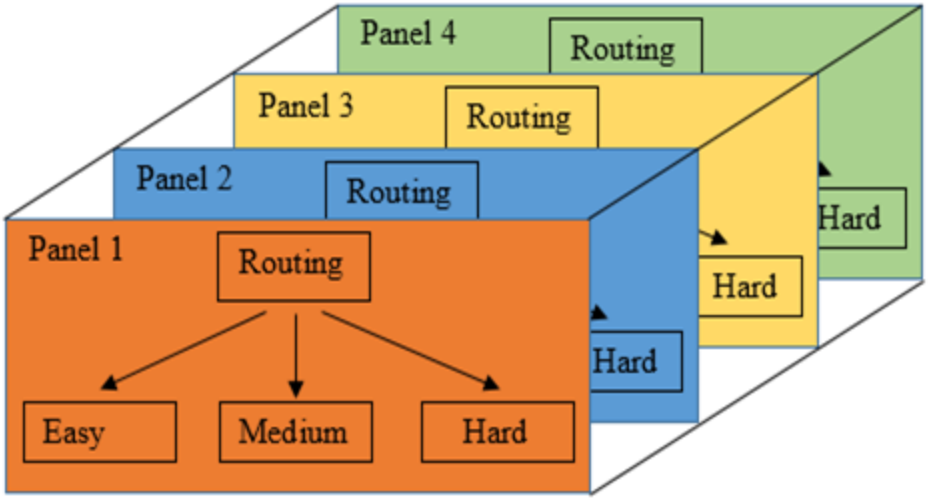

Figure 2-6. Illustration of multiple panels in ca-MST.

CHAPTER 3
METHODOLOGY

Design Overview

In this dissertation, one CAT design and two ca-MST panel designs were compared across several manipulated conditions. Two different ca-MST panel designs were simulated: the 1-3 and the 1-3-3 structure panel designs, with panels and modules constructed by integer programming. The common manipulated conditions across CAT and ca-MST were test length with two levels and number of controlled content area with five levels. All manipulated conditions within CAT and ca-MST were fully crossed with one another. This resulted in 2x5=10 CAT conditions (test length x content area), and 2x5x2=20 ca-MST conditions (test length x content area x ca-MST design), for 30 total conditions. For each condition 100 replications were performed. For better comparability, the following features were fixed across CAT and ca- ca-MST administrations: exposure rate, item bank, IRT model, and ability estimation method. Both varied and fixed conditions are detailed in following sections.

Fixed Conditions

The item parameters used in this dissertation were based on a real ACT math test as used in Luecht (1991) and Armstrong et al., (1996). The original item bank consisted of 480 multiple choice items from six content areas. The item parameters and number of items from each content in the original item bank were given in Table 3-1. In order to better compare the 30 conditions, and to avoid differences in content difficulty across the two, four, six and eight content area conditions, four additional item banks were generated (e.g., item bank 1, item bank 2, item bank 3 and item bank 4). These four different item banks were used in both CAT and ca-MST simulations, each for different content conditions. For the two content conditions, one relatively easy (Content 1) and one relatively hard content area (Content 6) from six available on the real

ACT test were selected (i.e., item bank 1). Under the four, six, and eight content conditions, the same sets of ACT items were used as in the two content condition, with multiple modules being developed from those sets. All conditions had an equal number of easy and hard content areas within each item bank to avoid content difficulty problems in the interpretations of the results.

All item banks had 480 items each with an equal number of items from each content area. This means that there were 240, 120, 80 and 60 items from each content area in item bank 1, 2, 3 and 4 respectively. Under the no content control condition, item bank 1 was used but content ID's were ignored. The item parameters and number of items for each content area in item bank 1 was given in Table 3-2. Again, the properties of items in other item pools are the same except the number of content areas and number of items. The total information functions for the four item banks are given in Figure 3-1. As expected and desired, the level of average item bank difficulty was very similar across the item banks.

Many conditions in this study were simulated for equivalency across CAT and ca-MST. These similarities allowed us to compare different content conditions not only within CAT and ca-MST designs, but also across the CAT and ca-MST designs. First, four thousand examinees were generated from a normal distribution, $N(0, 1)$. The theta values that represent examinees were re-generated for each replication, and used for CAT and two ca-MST simulations. Second, under a particular content area condition, ca-MST modules and panels for two different panel designs were built from the same item banks used for CAT simulations. Specifically, in the two, four, six and eight content ca-MST conditions, the modules and panels for both 1-3 and 1-3-3 panel designs were built from item bank 1, 2, 3 and 4, respectively. In both ca-MST and CAT, the 3PL IRT model (Birnbaum, 1968) was used (see Equation 2-3) to generate the item responses. Third, exposure rates were equal across the CAT and ca-MST simulations. The

maximum item level exposure rate in CAT simulations was 0.25. The four essentially parallel panels created in ca-MST designs also had an exposure rate of 0.25 is for a panel and for a module. In practice, only routing modules are seen by all examinees that are assigned to a panel, and the subsequent modules in each panel are seen by fewer examinees. Fourth, as the ability estimation method, the expected a posteriori (EAP) (Bock & Mislevy, 1982) with a prior distribution of $N(0, 1)$ was used for both interim and final ability estimates across the CAT and ca-MST. The whole simulation process for both CAT and ca-MST were completed in R version 2.1.1 (R Development Core Team, 2009-2015). Varied conditions in CAT and ca-MST simulations are detailed in the following sections.

CAT Simulations

There were two varying conditions in the CAT design; test length with two levels and number of controlled content area with five levels. The two different levels of test length were 24-item and 48-item length as used in similar simulation studies (Zenisky et al., 2010). This means that the test stopped after each test taker responds to the 24th and 48th item, respectively. The five levels of content area condition included zero (e.g., no content control), two, four, six and eight content area. No content control means that students did not necessarily receive pre-specified number of items from each content area. For example, in a biology test, while one student might receive eight botany items, eight ecology items, eight organ systems items, another student might receive less balanced items from these content areas. In the two content condition, the target proportions for content 1 and 2 were 50% and 50%, and the corresponding number of items were 12, 12 and 24, 24 for 24-item and 48-item test length conditions, respectively. In the four content condition, the target proportions for content 1, 2, 3 and 4 were 25% each, and the corresponding number of items were 6 and 12 for 24-item and 48-item test length conditions, respectively. In the six content condition, the target proportions for content 1, 2, 3, 4, 5 and 6

were 16.6% each, and the corresponding number of items were 4 and 8 for 24-item and 48-item test length conditions, respectively. In the eight content condition, the target proportions for content 1, 2, 3, 4, 5, 6, 7 and 8 were 12.5% each, and the corresponding number of items were 3 and 6 for 24-item and 48-item test length conditions, respectively. Under the no content control condition, students received a total of 24 and 48 items without ensuring number of items they received from each content area. The distribution of items across the contents areas and test lengths in CAT simulations are summarized in Table 3-3.

The literature showed that in terms of measurement accuracy, the constrained-CAT (CCAT; Kingsbury & Zara, 1989) method produced very similar results with other content control methods (Leung et al., 2003). So, the CCAT method was used for content control procedure as also used in many real CAT applications (Bergstrom & Lunz, 1999; Kingsbury, 1990; Zara, 1989). This method tracks the proportion of administered items for all contents. Then, the next item is selected from the content area that has the lowest proportion of administered items (e.g., largest discrepancy from the target proportion). The Sympson-Hetter (Sympson & Hetter, 1985) method with a fixed value of r_i= 0.25 was used for item exposure control.

Since the first theta estimate is calculated after responding to the first item, the initial theta is not known prior to the CAT administration. A typical approach to choosing the first item is to select an item of medium difficulty (i.e., $b = 0$), which was used in this study. The maximum information method (Brown & Weiss, 1977) was used as the item selection rule. This method selects the next item that provides the highest information about her current theta estimate by satisfying the content constraints.

Ca-MST Simulations

Two different ca-MST designs were built, the 1-3 structure design (e.g., two stage test) and the 1-3-3 structure design (e.g., three stage test). In any particular content area condition, all modules at any stage had the same number of items. This means that in the 1-3 panel design, there were 12 and 24 items per module under 24-item and 48-item test length conditions, respectively. In the 1-3-3 panel design, there were 8 and 16 items per module under 24-item and 48-item test length conditions, respectively. The distribution of items across the content areas in each module are given in Table 3-4 and Table 3-5 for 24-item and 48-item test length conditions, respectively. When the content was not controlled, the number of items in each module was the same under the stated conditions but the proportions across the content areas given in 3-4 and 3-5 were not necessarily met.

Previous research has shown that there are only slight differences between routing methods. In order to maximize the similarities between CAT and ca-MST, the maximum information method (Lord, 1980) was selected as the routing strategy because it can be similarly applied in both types of test administrations.

Test assembly. In both 1-3 and 1-3-3 ca-MST designs, four non-overlapping essentially parallel panels were generated from item bank 1, 2, 3 and 4. Creating multiple panels in ca-MST aimed to hold the maximum panel, module, and item exposure rates comparable to the CAT simulations. Panels and modules were created with IBM CPLEX (ILOG, Inc., 2006). This process was completed in two steps. First items were clustered into different modules, then modules were randomly assigned to the panels. The bottom-up strategy was used to create panels, which means that modules at the same difficulty level were exchangeable across the panels. As shown in Luecht (1998), the automated test assembly finds a solution to maximize the IRT information function at a fixed theta point (see Equation 2-4). Let denote $\theta 0$ is the fixed

theta point, and suppose we want a total of 24-item in the test. We first define a binary decision variable, xi, (e.g., xi=0 means item i is not selected from the item bank, xi=1 means item i is selected from the item bank). The information function we want maximize is;

$$I(\theta_0) = \sum_{i=1}^{N} I(\theta_0, \xi_i) x_i \tag{3-1}$$

where ξ_i represents the item parameters of item i (e.g., a, b, c parameters). Let's say we have two content areas (e.g., C_1 and C_2), and want to select an equal number of items from each content area. The automated test assembly is modeled to maximize

$$\sum_{i=1}^{N} I(\theta_0, \xi_i) x_i, \tag{3-2}$$

subject to

$$\sum_{i \in C1}^{N} x_i \geq 12 \tag{3-3}$$

$$\sum_{i \in C2}^{N} x_i \geq 12 \tag{3-4}$$

$$\sum_{i=1}^{N} x_i \geq 24 \tag{3-5}$$

$$x_i \in (0,1),\ i = 1, \ldots . N\ , \tag{3-6}$$

which put constraints on C_1, C_2, the total test length, and the range of decision variables, respectively. The test assembly models under other conditions (e.g., 48-item test length, six content area) can be modeled similarly. When content balancing was not controlled, the constraints on the contents were removed from the test assembly model.

In all conditions, the three fixed theta scores were chosen as θ1=-1, θ2=0, θ3=1, which represent the target information functions for easy, medium and hard modules, respectively. In both panel designs, the items in routing modules were chosen from medium difficulty items (e.g., items that maximize information function at theta point of 0). In the 1-3 panel design, there were

two medium modules, one easy and one hard module in each panel. In the 1-3-3 panel design, there were two easy, three medium and two hard modules in each panel. After modules were built, randomly assigned to the panels.

Evaluation Criteria

The results of the simulation were evaluated with two set of statistics: (a) overall results, (b) conditional results as evaluated in similar studies (see Han & Guo, 2014; Zenisky, 2004). For overall statistics, mean bias, root mean squared error (RMSE), and correlation between estimated and true theta ($\rho_{\hat{\theta}\theta}$) were computed from the simulation results as illustrated below.

Let *N* denote the total number of examinees, $\hat{\theta}_j$ the estimated theta score for person *j*, and θ_j the true theta score for person *j*. Mean bias was computed as

$$\bar{e} = \frac{\sum_{j=1}^{N}(\hat{\theta}_j - \theta_j)}{N}. \tag{3-7}$$

The RMSE was computed as

$$RMSE = \sqrt{\frac{\sum_{j=1}^{N}(\hat{\theta}_j - \theta_j)^2}{N}}. \tag{3-8}$$

The correlation between estimated and true theta was computed as

$$\rho_{\widehat{\theta_j},\theta_j} = \frac{cov\,(\hat{\theta}_j, \theta_j)}{\sigma_{\hat{\theta}_j}\sigma_{\theta_j}}, \tag{3-9}$$

where $\sigma_{\hat{\theta}_j}$and σ_{θ_j} are the standard errors for the estimated and true theta values, respectively. In any particular condition in CAT and ca-MST simulations, each overall statistic was calculated separately for each iteration across the 4,000 examinees, and then averaged across 100 replications. In addition, factorial ANOVA procedures were conducted to examine the statistically significant and moderately to large sized patterns among the simulation factors on these three outcomes. This was done separately for each outcome with the three study factors

(e.g., type of test administration- CAT vs 1-3 ca-MST vs 1-3-3 ca-MST, test length and number of content area) as fully crossed independent variables. For consistency, always the interaction of test administration model and content condition was plotted within each level test length.

For conditional result, conditional standard error of measurement (CSEM) was calculated. The CSEM was observed between θ =-2 and θ=2, and the width of the θ interval was 0.2. CSEM for a theta point was computed as

$$CSEM(\hat{\theta}_j) = \frac{1}{\sqrt{T(\hat{\theta}_j)}}. \tag{3-10}$$

It should be noted that the CSEM is the inverse of the square root of total test information at an estimated theta point.

Table 3-1. Item parameters of each content area in the original item bank

Content Area	*a*		*b*		*c*	
(Number of items)	Mean	SD	Mean	SD	Mean	SD
Content 1 (*n*=48)	1.015	.292	-.485	.465	.154	.044
Content 2 (*n*=168)	.911	.322	.131	.977	.160	.054
Content 3 (*n*=24)	1.028	.328	.811	.778	.173	.059
Content 4 (*n*=96)	1.120	.419	.689	.655	.167	.058
Content 5 (*n*=96)	1.037	.356	.527	.650	.151	.062
Content 6 (*n*=48)	.911	.312	.475	.828	.163	.058

Table 3-2. Item parameters of each content area in item bank 1

Content Area	*a*		*b*		*c*	
(Number of items)	Mean	SD	Mean	SD	Mean	SD
Content 1 (*n*=300)	1.015	.292	-.485	.465	.154	.044
Content 2 (*n*=300)	.911	.312	.475	.828	.163	.058

Table 3-3. Distribution of items across the contents areas and test lengths in CAT

Test Length	Content Condition	Content Areas	Target Proportions	Corresponding Number of Items
24-item CAT	Two Content	C_1,C_2	50%,50%	12,12
	Four Content	C_1,C_2,C_3,C_4	25%,25%,25%25%	6,6,6,6
	Six Content	C_1,C_2,C_3,C_4,C_5,C_6	16.6%,16.6%,16.6%, 16.6%,16.6%,16.6%	4,4,4,4,4,4
	Eight Content	$C_1,C_2,C_3,C_4,C_5,C_6,C_7,C_8$	12.5%,12.5%,12.5%,12.5%, 12.5%,12.5%,12.5%,12.5%	3,3,3,3,3,3,3,3
48-item CAT	Two Content	C_1,C_2	50%,50%	24,24
	Four Content	C_1,C_2,C_3,C_4	25%,25%,25%25%	12,12,12,12
	Six Content	C_1,C_2,C_3,C_4,C_5,C_6	16.6%,16.6%,16.6%, 16.6%,16.6%,16.6%	8,8,8,8,8,8
	Eight Content	$C_1,C_2,C_3,C_4,C_5,C_6,C_7,C_8$	12.5%,12.5%,12.5%,12.5%, 12.5%,12.5%,12.5%,12.5%	6,6,6,6,6,6,6,6

Table 3-4. The distribution of items in modules in ca-MST across the content areas under 24-item test length

	Content Condition	Content Area	Routing Modules	Stage 2 Modules	Stage 3 Modules	Total
1-3 Panel Design	Two Content	C_1,C_2	6,6	6,6	-	24
	Four Content	C_1,C_2,C_3,C_4	3,3,3,3	3,3,3,3	-	24
	Six Content	C_1,C_2,C_3,C_4,C_5,C_6	2,2,2,2,2,2	2,2,2,2,2,2	-	24
	Eight Content	$C_1,C_2,C_3,C_4,C_5,C_6,C_7,C_8$	2,2,2,2,1,1,1,1	1,1,1,1,2,2,2,2	-	24
1-3-3 Panel Design	Two Content	C_1,C_2	4,4	4,4	4,4	24
	Four Content	C_1,C_2,C_3,C_4	2,2,2,2	2,2,2,2	2,2,2,2	24
	Six Content	C_1,C_2,C_3,C_4,C_5,C_6	2,2,1,1,1,1	1,1,2,2,1,1	1,1,1,1,2,2	24
	Eight Content	$C_1,C_2,C_3,C_4,C_5,C_6,C_7,C_8$	1,1,1,1,1,1,1,1	1,1,1,1,1,1,1,1	1,1,1,1,1,1,1,1	24

Table 3-5. The distribution of items in modules in ca-MSTs across the content areas in under 48-item test length

	Content Condition	Content Area	Routing Modules	Stage 2 Modules	Stage 3 Modules	Total
1-3 Panel Design	Two Content	C_1,C_2	12,12	12,12	-	48
	Four Content	C_1,C_2,C_3,C_4	6,6,6,6	6,6,6,6	-	48
	Six Content	C_1,C_2,C_3,C_4,C_5,C_6	4,4,4,4,4,4	4,4,4,4,4,4	-	48
	Eight Content	$C_1,C_2,C_3,C_4,C_5,C_6,C_7,C_8$	3,3,3,3,3,3,3,3	3,3,3,3,3,3,3,3	-	48
1-3-3 Panel Design	Two Content	C_1,C_2	8,8	8,8	8,8	48
	Four Content	C_1,C_2,C_3,C_4	4,4,4,4	4,4,4,4	4,4,4,4	48
	Six Content	C_1,C_2,C_3,C_4,C_5,C_6	4,4,2,2,2,2	2,2,4,4,2,2	2,2,2,2,4,4	48
	Eight Content	$C_1,C_2,C_3,C_4,C_5,C_6,C_7,C_8$	2,2,2,2,2,2,2,2	2,2,2,2,2,2,2,2	2,2,2,2,2,2,2,2	48

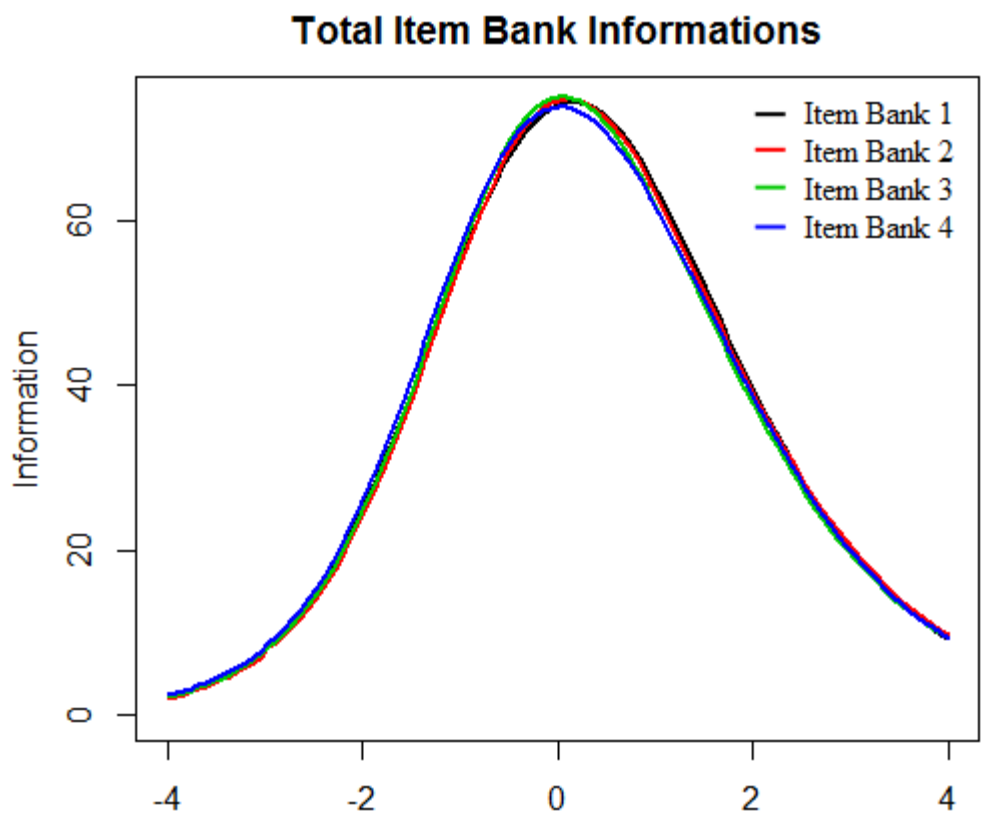

Figure 3-1. Total information functions for four item banks.

CHAPTER 4
RESULTS

In this chapter, the results of this study are presented in two sections. The first section describes the overall findings (e.g., mean bias, RMSE, correlations between the estimated and true theta values), and the second section describes the conditional results (e.g., CSEM, conditional mean bias).

Overall Results

Mean Bias

The results of mean bias across the conditions were presented in Table 4-1. To assess for statistically significant patterns, a factorial ANOVA was conducted with mean bias as the outcome and the three study factors (e.g., test administration model, test length, number of content area) as the independent variables. The results for factorial ANOVA were presented in Table 4-2.

The main effect of test administration explained the largest proportion of mean bias variance (η^2=.17), controlling for all other factors. The interactions or other main effects were either non-significant or explained very small proportion of variance. A graphical depiction of the main effect of test administration within each level of test length on the mean bias is displayed in Figure 4-1. The main finding was that regardless of the number of content area and test length, CAT produced lower amount of mean bias than the two ca-MTS's, which caused the main effect of test administration. Another finding was that the 1-3 and the 1-3-3 panel designs resulted in very similar mean bias. However, the test length did not impact the mean bias across the conditions.

Root Mean Square Error

The results of RMSE across the conditions were presented in Table 4-3. To assess for statistically significant patterns, a factorial ANOVA was conducted with RMSE as the outcome and the three study factors (e.g., test administration model, test length, number of content area) as the independent variables. The results for factorial ANOVA were presented in Table 4-4.

The interaction of test administration model and test length explained a meaningful proportion of RMSE variance (η^2=.05), as did the main effect of test length (η^2=.18). A graphical depiction of the interaction of test administration model and test length within each level of number of content area on the RMSE is displayed in Figure 4-2. The main finding was that regardless of the number of content area and test administration model, as the test length increased, the amount of RMSE decreased. Also, the decrease in RMSE associated with the increase in test length was almost always more obvious for CAT, which caused the significant two way interaction of test administration and test length. The varying ca-MST panel designs did not impact the amount of RMSE within each level of test length as well as the number of content area.

Correlation

The results of correlation between the true and estimated theta values across the conditions were presented in Table 4-5. To assess for statistically significant patterns, a factorial ANOVA was conducted with correlation as the outcome and the three study factors (e.g., test administration model, test length, number of content area) as the independent variables. The results for factorial ANOVA were presented in Table 4-6.

The main effect of test administration and test length explained the largest proportion of correlation variance (η^2=.35), controlling for all other factors. The interactions or other main effects were either non-significant or explained very small proportion of variance. A graphical

depiction of the main effect of test administration model within each level of test length on the correlations is displayed in Figure 4-3. The main finding was that regardless of the number of content area and test administration model, as the test length increased, correlation between the true and estimated thetas increased, which caused significant main effect of test length. Another main finding was that regardless of the number of content area and test length, the correlations were always lower under CAT, which caused significant main effect of test administration model. However, the number of content area did not impact the correlations. The varying ca-MST panel designs did not impact the correlations within each level of test length as well.

Conditional

Conditional standard error of measurement. Results of standard error of measurement conditioned on the estimated theta values across the two different test lengths and three different are displayed in Figure 4-4. First finding was that standard error of measurements were always lowest around $\theta=0$ point within any condition. Second finding was that as the test length increased standard error of measurements throughout the estimated theta points decreased for all test administration models. Third finding was that standard error of measurements for the two ca-MST conditions were more stable (more consistent) than CAT, regardless of the test length. Fourth finding was that even though the fluctuations were greater for CAT, CAT resulted in lower standard error of measurements throughout the estimated theta points. However, the number of content area did not substantially impact the standard error of measurements within a particular condition, and the interpretations were always similar across the number of content area conditions.

Table 4-1. Results of mean bias across the conditions

Design	Test Length	No Control	2 Content	4 Content	6 Content	8 Content
1-3 ca-MST	24-item	0.07	0.07	0.05	0.05	0.06
1-3 ca-MST	48-item	0.06	0.07	0.06	0.07	0.07
1-3-3 ca-MST	24-item	0.06	0.05	0.06	0.05	0.07
1-3-3 ca-MST	48-item	0.08	0.05	0.05	0.04	0.06
CAT	24-item	0.00	0.00	0.00	0.00	-0.00
CAT	48-item	-0.00	0.00	-0.00	0.00	0.00

Table 4-2. Factorial ANOVA findings when dependent variable was mean bias

Dependent Variable	*p*	η^2
Test Administration x Number of Content x Test Length	.15	.00
Test Administration x Number of Content	.04	.00
Test Administration x Test Length	.22	.00
Number of Content x Test Length	.25	.00
Test Administration	.00	.17
Number of Content	.05	.00
Test Length	.61	.00

Table 4-3. Results of RMSE across the conditions

Design	Test Length	No Control	2 Content	4 Content	6 Content	8 Content
1-3 ca-MST	24-item	0.33	0.33	0.35	0.33	0.34
1-3 ca-MST	48-item	0.30	0.30	0.31	0.31	0.30
1-3-3 ca-MST	24-item	0.32	0.32	0.33	0.34	0.34
1-3-3 ca-MST	48-item	0.30	0.31	0.31	0.31	0.31
CAT	24-item	0.35	0.36	0.36	0.37	0.36
CAT	48-item	0.28	0.29	0.29	0.28	0.29

Table 4-4. Factorial ANOVA findings when dependent variable was RMSE

Dependent Variable	*p*	η^2
Test Administration x Number of Content x Test Length	.04	.00
Test Administration x Number of Content	.48	.00
Test Administration x Test Length	.00	.05
Number of Content x Test Length	.38	.00
Test Administration	.06	.00
Number of Content	.00	.02
Test Length	.00	.18

Table 4-5. Correlation coefficients between true and estimated thetas across the conditions

Design	Test Length	No Control	2 Content	4 Content	6 Content	8 Content
1-3 ca-MST	24-item	0.95	0.95	0.95	0.95	0.95
1-3 ca-MST	48-item	0.97	0.97	0.97	0.97	0.97
1-3-3 ca-MST	24-item	0.96	0.95	0.95	0.95	0.95
1-3-3 ca-MST	48-item	0.97	0.97	0.97	0.97	0.97
CAT	24-item	0.93	0.93	0.93	0.92	0.93
CAT	48-item	0.95	0.95	0.95	0.95	0.95

Table 4-6. Factorial ANOVA findings when dependent variable was correlation

Dependent Variable	*p*	η^2
Test Administration x Number of Content x Test Length	.12	.00
Test Administration x Number of Content	.04	.00
Test Administration x Test Length	.00	.02
Number of Content x Test Length	.00	.00
Test Administration	.00	.35
Number of Content	.00	.00
Test Length	.00	.35

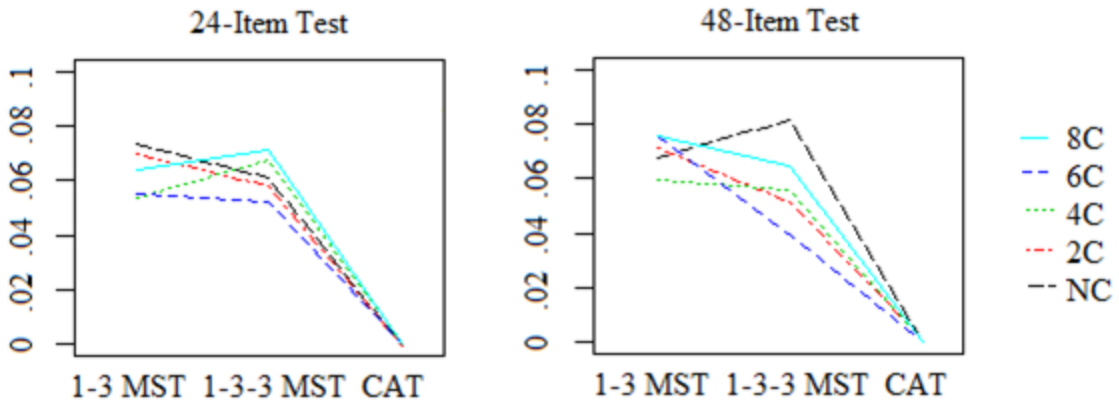

Figure 4-1. Main effect of test administration model on mean bias within levels of test length.

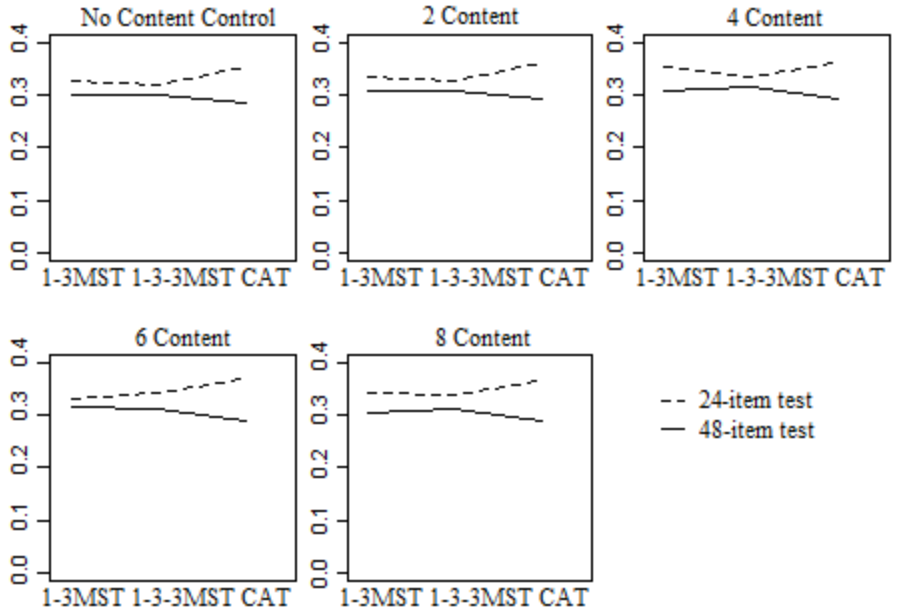

Figure 4-2. Interaction of test administration model and test length on RMSE within levels of number of content.

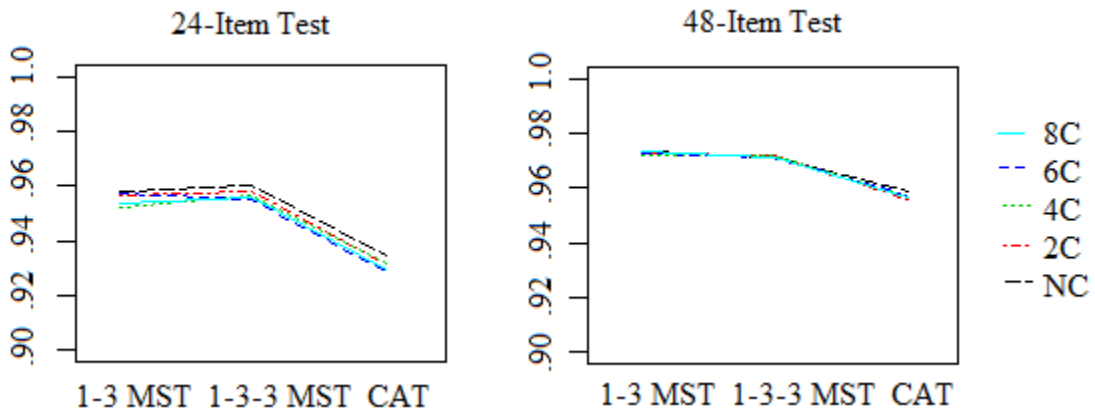

Figure 4-3. Main effect of test administration model on correlation within levels of test length.

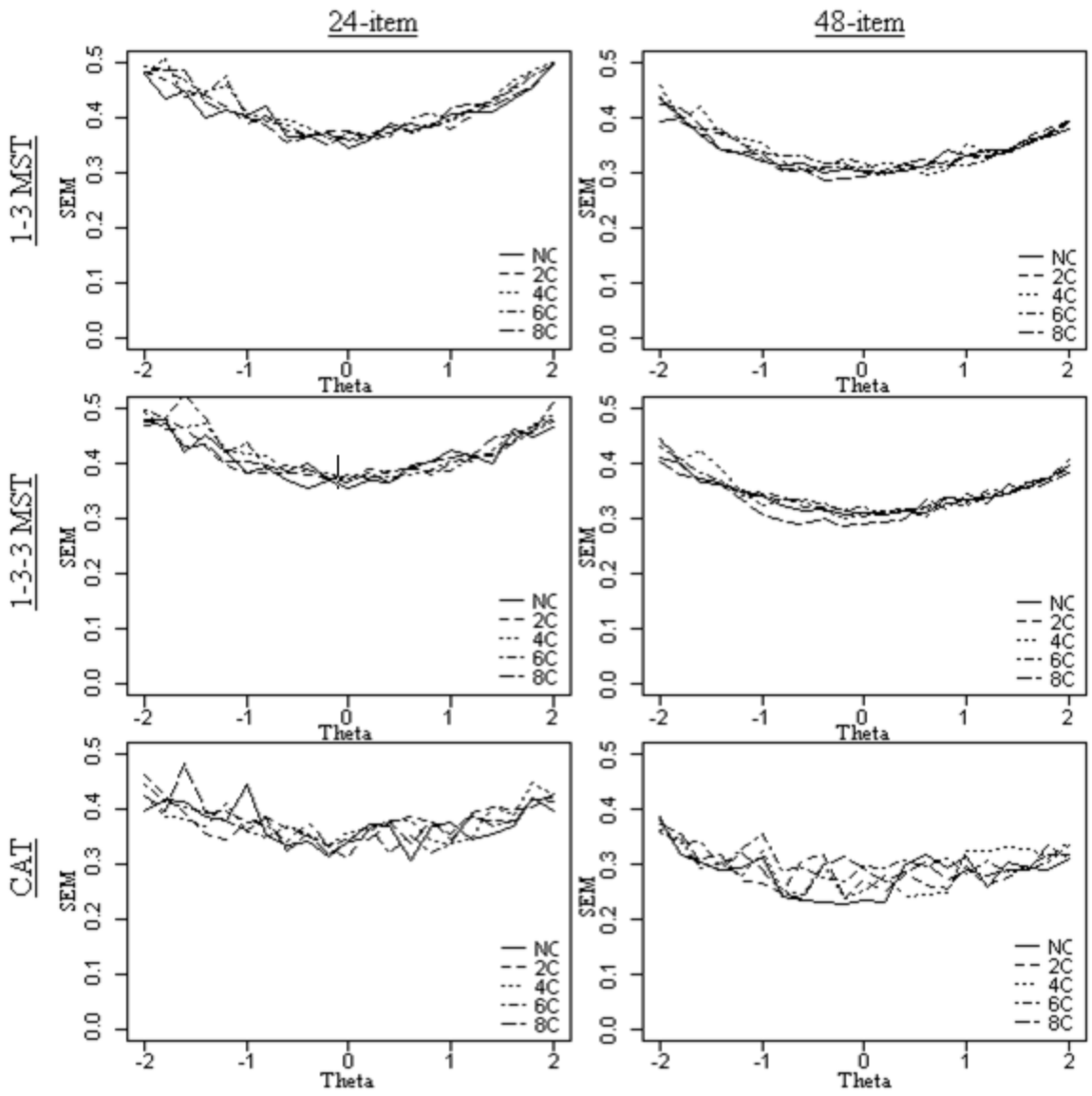

Figure 4-4. Conditional standard error of measurements

CHAPTER 5
DISCUSSION AND LIMITATIONS

The main purpose of this study was to explore the precision of test outcomes across computerized adaptive testing and computerized multistage testing when the number of different content areas was varied across the different test lengths. It was important to examine this because content balancing and content alignment is a requirement for validity of score-based inferences (Messick, 1989; Wise, & Kingsbury, 2015). In real applications, item pools most often have items from multiple content areas, and dealing with content control might not be easy in adaptive testing (Wise & Kingsbury, 2015). However, the consequences of not ensuring content balancing can have potential negative effects on the test use and score interpretations. Hence, this study added to the literature on content control in adaptive testing and, more specifically, aimed to provide guidelines about the relative strengths and weaknesses of ca-MST as compared to CAT with respect to content control.

It is important to note that the content itself cannot be generated in a simulation study. Rather, item parameters are generated and used to represent different content areas. In this study, we defined a content area as a group of items, which belong to a specific sub-curriculum of the test, such as fractions, algebra or calculus. However, when generating these content areas, we used a particular range of item parameters to represent such sub-curriculum sections of the test. This process has limitations, but is similar to approaches in other similar simulation studies (see Armstrong et al., 1996; Luecht, 1991).

The results showed that in terms of mean bias, CAT produced slightly better results than two ca-MSTs. This was the only meaningful finding, and other study factors did not have substantial impact on the mean bias (see Table 4-2 and Figure 4-1). In terms of RMSE, only test length had meaningful impact on the outcome (see Table 4-4); increasing test length improved

the outcome (see Figure 4-2). In terms of correlations between the true and estimated theta values, both test length and type of test administration played important role on the outcome (see Table 4-6). The two ca-MSTs produced very comparable results and outperformed CAT. Furthermore, increasing test length improved the outcome of correlation (see Figure 4-1). However, it is important note that low correlation does not mean results are poorer for CAT. As it is seen in the Figure 4-4, the fluctuations in the conditional standard error of measurements were greater for CAT than that for the two ca-MSTs. This was the reason behind the lower correlations under CAT. Even if the two ca-MSTs provided more stable standard error of measurements across the different theta values (much flatter), CAT produced lower standard error of measurements than the two ca-MSTs (see Figure 4-3 and Figure 4-4). The effect of test length was more obvious when the standard error of measurements were plotted against the theta values (see Figure 4-4).

This study did not find any evidence of the effect of content area on both CAT and ca-MSTs. Increasing the number of controlled content areas or having no control over content did not meaningfully affect any of the study outcomes. All three test administration models were able to find items which provide the most information from different content areas, regardless of the number of content conditions. For practitioners and researchers, this is an indication that the studied CAT and ca-MST methods are not unduly influenced by the number of content areas one might have on a test. Thus, this is a positive finding for the practitioners and researchers as CAT or ca-MST designers do not need to worry about the number of content areas.

This study found that there was no meaningful difference on the outcomes between the two ca-MST panel designs, but CAT outperformed the ca-MSTs under many conditions. This study however does not argue against the ca-MSTs. Even though these were outside of the

primary purpose of this study, the ca-MSTs have several practical advantages in real applications. First, due to the feature of test assembly, even if the measurement precision is lower in ca-MSTs then CAT, all other parameters being equal, the ca-MST allows greater control over test design and content. The ca-MST designers can determine item and content order within modules in ca-MST. There is however no expert control over relative item order in CAT. Second, since items in the modules are placed by the test developer prior to the administration, ca-MST allows for strict adherence to content specification, no matter how complex it is (Yan et al., 2014). However, in CAT, content misspecifications are more likely to occur (see Leung et al., 2003). Third, ca-MST allows adding other constraints on placing items into the modules, such as item length and format. This means that the length of items in different modules and panels can easily be controlled in ca-MST whereas not traditionally controlled. Fourth, ca-MST allowed higher percentage of item pool use over CAT. In 1-3 ca-MST condition, item pool usage rates were 40% and 80% under 24 and 48-item test lengths, respectively. In 1-3-3 ca-MST condition, item pool usage rates were 46.6% and 93.3% under 24 and 48-item test lengths, respectively. However, pool usage rates for CAT were about 36% and 64% under 24 and 48-item test lengths, respectively. Fifth issue is item retirement for cost. As stated before, the cost for a single item varies from $1,500 to $2,500 (Rudner, 2009). Besides the high pool usage rate, having less number of retired items is desired because researchers and practitioners do not want to throw many items away after each administration. In ca-MST only the items in the routing modules reach pre-specified maximum control rate, and are then retired for use. In 1-3 ca-MST, the number of items with maximum exposure rates were 48 (e.g., 12 items per a routing module in a panel x 4 different panels) and 96 (e.g., 24 items per a routing module in a panel x 4 different panels) under 24 and 48-item length conditions, respectively. In 1-3-3 ca-MST, the

number of items with maximum exposure rates were 32 (e.g., 8 items per a routing module in a panel x 4 different panels) and 64 (e.g., 16 items per a routing module in a panel x 4 different panels) under 24 and 48-item length conditions, respectively. In CAT, the number of items with maximum exposure rates were 47 and 105 under 24 and 48-item length conditions, respectively. Apparently, the 1-3-3 ca-MST resulted in the lowest number of items with maximum exposure rate. As a result, fewer items were retired for use. In fact, this number could be further reduced by placing less items into the routing modules, and thus more items can be saved for future administrations. This characteristic is another advantage of ca-MST over CAT. A further research can investigate item pool utilization under varying conditions such as item bank size, quality of item bank, different ca-MST panel designs and varying level of number of items in the routing module. It is necessary to note that, for a fair comparison, overuse of some items in the bank was more likely due to the nature of maximum fisher information item selection method in CAT. This method always selects best items multiple time for different examinees until an item reaches the exposure rate, which reduces item pool usage and increases number of retired items. This could be controlled with a different item selection method.

This study argues that although ca-MST allows greater flexibility over the content control, there was a reduction in measurement efficiency. This is not surprising because of the lack of adaptation points. In ca-MST, the number of adaptation points is associated with the number of stages (e.g., one minus number of stages), whereas it is associated with number of items in CAT (e.g., one minus number of items). For example, in 1-3 and 1-3-3 ca-MST panel designs, there was one and two adaptation points regardless of the test length, respectively. In CAT, there were 23 and 47 adaptation points under 24 and 48-item test length conditions, respectively.

In this study, we set the test length equal across the CAT and ca-MST simulations and investigated measurement accuracy as an outcome. However, measurement accuracy outcomes (e.g., standard error of measurement) are often used as stopping rules in CAT. In this study, ca-MST conditions were associated with a reduction in measurement accuracy as compared to CAT. However, if measurement accuracy were to be used as a stopping rule, we expect the outcomes to be quite different. It seems quite plausible that with this different type of stopping rule, CAT and ca-MST would have the same measurement accuracy outcomes but would differ in the test lengths needed to obtain those outcomes. In summary, the restriction of equal test length and the choice of stopping rule in this study had large impacts on the outcomes. Future studies may change and/or vary these conditions to explore outcomes across both test administration models.

In order to give similar advantage to both CAT and ca-MSTs, we intentionally used maximum Fisher information method as an item selection and a routing method, for CAT and ca-MST, respectively. However, this study can be improved with running the same simulation by adopting other item selections and routing methods. Additionally, in order to avoid content difficulty confounding, we systematically chose two content areas as one easy and one hard content area. However, in real applications this will not likely happen and test content might have wider and non-systematic ranges of contents areas in terms of difficulty (e.g., a few easy, several medium, and many hard content areas). We recommend the studied conditions to be tested with a real test administration that does not have systematic difficulty differences across content areas. In a particular number of content conditions, there were equal number of items from different content areas. This study should also be replicated with unequal content distributions. Furthermore, even though the main purpose of this study was to examine the effect of the content control across adaptive testing, a further study should include linear test, and use it

as a baseline condition when comparing with CAT and ca-MST. A further study can also be replicated with different exposure rates and item bank sizes. Additionally, it is known that items form a content area do not only measure that content, but also measure other content domains with possibly small loadings. In other words, the assumption of unidimensionality is rarely met in real applications (Lau, 1997; Reckase, 1979).

LIST OF REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association

Angoff, W. H., & Huddleston, E. M. (1958). *The multi-level experiment: a study of a two-level test system for the College Board Scholastic Aptitude Test. Princeton, New Jersey, Educational Testing Service*. Statistical Report SR-58-21.

Armstrong, R. D. & L. Roussos. (2005). A method to determine targets for multi-stage adaptive tests. (Research Report 02-07). Newtown, PA: Law School Admissions Council.

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

Armstrong, R. D., Jones, D. H., Li, X., & Wu, L. (1996). A study of a network-flow algorithm and a noncorrecting algorithm for test assembly. *Applied Psychological Measurement*, *20*(1), 89-98.

Babcock, B. & Weiss, D. J. (2009). Termination criteria in computerized adaptive tests: Variable length CATs are not biased. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing*.

Baker, F. (1992). *Item response theory*. New York, NY: Markel Dekker, INC.

Barrada, J. R., Olea, J., Ponsoda, V., and Abad, F. J. (2008). Incorporating randomness to the Fisher information for improving item exposure control in CATS. *British Journal of Mathematical and Statistical Psychology*, *61*, 493-513.

Becker, K. A., & Bergstrom, B. A. (2013). Test administration models. *Practical Assessment, Research & Evaluation*, *18*(14), 7.

Belov, D. I., & Armstrong, R. D. (2005). Monte Carlo test assembly for item pool analysis and extension. *Applied Psychological Measurement*, *29*(4), 239-261.

Bergstrom, B. A., & Lunz, M. E. (1999). CAT for certification and licensure.*Innovations in computerized assessment*, 67-91.

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

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied psychological measurement*, *6*(4), 431-444.

Boyle, M., Jones, P., & Matthews-Lopez, J. (Feb, 2012). Replacing an Adaptive Decision-Making Test with a Pallet of Automatically Assembled Fixed Forms. Breakout Session at Association of Test Publishers Conference, Palm Springs, CA.

Bridgeman, B. (2012). A Simple Answer to a Simple Question on Changing Answers. *Journal of Educational Measurement*, *49*(4), 467-468.

Brown, J. M., & Weiss, D. J. (1977). *An adaptive testing strategy for achievement test batteries* (Research Rep. No. 77-6). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, Computerized Adaptive Testing Laboratory.

Chang, H. H., & Ying, Z. (1999). A-stratified multistage computerized adaptive testing. *Applied Psychological Measurement*, *23*(3), 211-222.

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

Chang, S.W., & Ansley, T. N. (2003). A comparative study of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement*, *40*, 71-103.

Chen, Y., & Ankenman, R. D. (2004). Effects of practical constraints on item selection rules at the early stages of computerized adaptive testing. *Journal of Educational Measurement*, *41*(2), 149-174.

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

Choi, S. W., Grady, M. W., & Dodd, B. G. (2010). A New Stopping Rule for Computerized Adaptive Testing. *Educational and Psychological Measurement*, *70*(6), 1–17.

Chuah, S. C., Drasgow, F., & Luecht, R. (2006). How big is big enough? Sample size requirements for CAST item parameter estimation. *Applied Measurement in Education*, *19*(3), 241-255.

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

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement*, (2nd ed., pp. 443–507). Washington DC: American Council on Education.

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

Crotts, K. M., Zenisky, A. L., & Sireci, S. G. (2012, April). *Estimating measurement precision in reduced-length multistage-adaptive testing.* Paper presented at the meeting of the National Council on Measurement in Education, Vancouver, BC, Canada.

Davey, T., & Y.H. Lee. (2011). Potential impact of context effects on the scoring and equating of the multistage GRE Revised General Test. (GRE Board Research Report 08-01). Princeton, NJ: Educational Testing Service.

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

Dubois, P. H. (1970). *A history of psychological testing*. Boston: Allyn & Bacon

Eggen, T. J. (1999). Item selection in adaptive testing with the sequential probability ratio test. *Applied Psychological Measurement*, *23*(3), 249-261.

Eignor, D. R., Stocking, M. L., Way, W. D., & Steffen, M. (1993). Case studies in computer adaptive test design through simulation. *ETS Research Report Series*, *1993*(2).

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

Flaugher, R. (2000). Item Pools. In Wainer, H. (Ed.), Computerized adaptive testing: A primer (2nd ed.). Mahwah, NH: Lawrence Erlbaum Associates.

Georgiadou, E. G., Triantafillou, E., & Economides, A. A. (2007). A review of item exposure control strategies for computerized adaptive testing developed from 1983 to 2005. *The Journal of Technology, Learning and Assessment*, *5*(8).

Gibson, W. M., & Weiner, J. A. (1998). Generating random parallel test forms using CTT in computer-based environment. Journal of Educational Measurement, *35*(4), 297-310.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, *12*(3), 3847.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications* (Vol. 7). Springer Science & Business Media.

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

Han, K. C. T., & Guo, F. (2014). Multistage Testing by Shaping Modules on the Fly. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized Multistage Testing: Theory and Applications* (pp. 119-133). Chapman and Hall/CRC.

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

Huang, S. X. (1996, January). A content-balanced adaptive testing algorithm for computer-based training systems. In *Intelligent Tutoring Systems* (pp. 306-314). Springer Berlin Heidelberg.

ILOG. (2006). ILOG CPLEX 10.0 [User's manual]. Paris, France: ILOG SA.

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

Kane, M. (2010). Validity and fairness. *Language Testing*, *27*(2), 177.

Keller, L. A. (2000). Ability estimation procedures in computerized adaptive testing. *USA: American Institute of Certified Public Accountants-AICPA Research Concortium-Examination Teams*.

Keng, L. & Dodd, B.G. (2009, April). *A comparison of the performance of testlet based computer adaptive tests and multistage tests*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Keng, L. (2008). *A comparison of the performance of testlet-based computer adaptive tests and multistage tests* (Order No. 3315089).

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

Kingsbury, G. G., & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education*, *2*(4), 359-375.

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

Lau, C. M. A. (1997). Robustness of a Unidimensional Computerized Mastery Testing Procedure With Multidimensional Testing Data. Dissertation Abstract International, *57*(7).

Leung, C.K., Chang, H.H., & Hau, K.T. (2000, April). *Content balancing in stratified computerized adaptive testing designs*. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Leung, C.K., Chang, H.H., & Hau, K.T. (2002). Item selection in computerized adaptive testing: Improving the a-Stratified design with the Sympson-Hetter algorithm. *Applied Psychological Measurement*, *26*(4), 376–392.

Leung, C.K., Chang, H.H., & Hau, K.T. (2003). Computerized adaptive testing: A comparison of three content balancing methods. *Journal of Technology, Learning, and Assessment*, *2*(5).

Lewis, C., & Sheehan, K. (1990). Using Bayesian decision theory to design a computerized mastery test. *Applied Psychological Measurement*, *14*(4), 367-386.

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

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

Lord, F. M. (1974). Practical methods for redesigning a homogeneous test, also for designing a multilevel test. Educational Testing Service RB-74–30.

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

Lord, F. M. (1986). Maximum likelihood and Bayesian parameter estimation in item response theory. *Journal of Educational Measurem*ent, *23*, 157–162.

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

Luecht, R. M. & Sireci, S. G. (2011). A review of models for computer-based testing. Research Report RR-2011-12. New York: The College Board.

Luecht, R. M. (1991). [American College Testing Program : Experimental item pool parameters.] Unpublished raw data.

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

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

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

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

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

Luecht, R. M., de Champlain, A., & Nungester, R. J. (1998). Maintaining content validity in computerized adaptive testing. *Advances in Health Sciences Education*, *3*(1), 29-41.

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

McBride, J. R., & Martin, J. T. (1983). Reliability and validity of adaptive ability tests in a military setting. *New horizons in testing*, 223-236.

Mead, A. D. (2006). An introduction to multistage testing. *Applied Measurement in Education*, *19*(3), 185-187.

Messick S. (1989) in *Educational Measurement*, Validity, ed Linn RL (American Council on Education/Macmillan, New York), 3rd ed, p. 13–103.

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.

Parshall, C. G., Spray, J., Kalohn, J., Davey, T. (2002). Practical considerations in computer-based testing. New York: Springer-Verlag.

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

Patsula, L. N. (1999). *A comparison of computerized adaptive testing and multistage testing* (Order No. 9950199). Available from ProQuest Dissertations & Theses Global. (304514969)

R Development Core Team. (2013). *R: A language and environment for statistical computing, reference index* (Version 2.2.1). Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org

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

Reckase, M. D. (1979). Unifactor Latent Trait Models Applied to Multifactor Tests: Results and Implications. *Journal of Educational Statistics*, 4(*3*), 207-230.

Reckase, M. D. (1981). *The formation of homogeneous item sets when guessing is a factor in item responses* Retrieved from http://search.proquest.com/docview/87952861?accountid=10920

Reese, L. M., & Schnipke, D. L (1999). *A comparison of] testlet-based test designs for computerized adaptive testing. Law school admission council computerized testing report. LSAC research report series.* (No. LSAC-R-97-01).

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

Rudner, L. M. (2009). Implementing the graduate management admission test computerized adaptive test. In *Elements of adaptive testing* (pp. 151-165). Springer New York.

Schnipke, D. L., & Reese, L. M. (1999). A Comparison [of] Testlet-Based Test Designs for Computerized Adaptive Testing. Law School Admission Council Computerized Testing Report. LSAC Research Report Series.

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika,61*(2), 331-354.

Segall, D. O. (2005). Computerized adaptive testing. *Encyclopedia of social measurement, 1*, 429-438.

Sireci, S. G. (2007). On validity theory and test validation. *Educational Researcher, 36*(8), 477-481.

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

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

Stocking, M.L., Smith, R., & Swanson, L. (2000). An investigation of approaches to computerizing the GRE subject tests. Research Report 00-04. Princeton, NJ: Educational Testing Service.

Sympson, J. B., & Hetter, R. D. (1985). Controlling item-exposure rates in computerized adaptive testing. In *Proceedings of the 27th annual meeting of the Military Testing Association* (pp. 973-977).

Thissen, D., & Mislevy, R. J. (2000). Testing algorithms. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (2nd ed., pp. 101-133). Hillsdale, NJ: Lawrence Erlbaum.

Thompson, N. A. (2008). A Proposed Framework of Test Administration Methods. *Journal of Applied Testing Technology, 9*(5), 1-17.

van der Linden, W. J., & Glas, C. A. (Eds.). (2000). *Computerized adaptive testing: Theory and practice*. Dordrecht: Kluwer Academic.

van der Linden, W. J., & Pashley, P. J. (2000). Item selection and ability estimation in adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), Computerized adaptive testing: Theory and practice (pp. 1-25). Norwell, MA: Kluwer.

van der Linden, W.J. & Chang, H.H. (2005, August). Implementing content constraints in alpha-stratified adaptive testing using a shadow test approach. Law School Admission Council, Computerized Testing Report (01-09).

van der Linden, W.J., Jeon, M., & Ferrara, S. (2011). A paradox in the study of the benefits of test-item review. *Journal of Educational Measurement, 48*(4), 380-398.

Veerkamp, W. J., & Berger, M. P. (1994). Some New Item Selection Criteria for Adaptive Testing. Research Report 94-6. Enschede: University of Twente, Faculty of Educational Science and Technology.

Veldkamp, B. P., & van der Linden, W. J. (2002). Multidimensional adaptive testing with constraints on test content. *Psychometrika*, *67*(4), 575-588.

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

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., & Mislevy, R. J. (2000). *Computerized adaptive testing: A primer*. Routledge.

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

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

Ward, W. C. (1988). The College Board Computerized Placement Tests: An Application of Computerized Adaptive Testing. *Machine-Mediated Learning*, *2*, 271-82.

Weiss, D. J. (1973). *The stratified adaptive computerized ability test* (Research Report 73-3). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Weiss, D. J., & Betz, N. E. (1974). *Simulation Studies of Two-Stage Ability Testing* (No. RR-74-4). Minnesota Univ Minneapolis Dept of Psychology.

Weiss, D. J., & Kingsbury, G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, *21*(4), 361-375.

Weissman, A., Belov, D.I., & Armstrong, R.D. (2007). Information-based versus number-correct routing in multistage classification tests. (Research Report RR-07-05). Newtown, PA: Law School Admissions Council.

Wise, S. L., Kingsbury, G. G., & Webb, N. L. (2015). Evaluating Content Alignment in Computerized Adaptive Testing. *Educational Measurement: Issues and Practice*, *34*(4), 41-48.

Yan, D., von Davier, A. A., & Lewis, C. (Eds.). (2014). *Computerized multistage testing: Theory and applications*. CRC Press.

Yi, Q., Wang, T., & Ban, J.C. (2001). Effects of scale transformation and test-termination rule on the precision of ability estimation in computerized adaptive testing. *Journal of Educational Measurement*, *38*, 267-292.

Yi, Q., Zhang, J., & Chang, H. H. (2006). Assessing CAT Test Security Severity. *Applied psychological measurement*, *30*(1), 62-63.

Zara, A. R. (1989). A research proposal for field testing CAT for nursing licensure examinations. *Delegate Assembly Book of Reports*.

Zenisky, A. L. (2004). *Evaluating the effects of several multi-stage testing design variables on selected psychometric outcomes for certification and licensure assessment* (Order No. 3136800).

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

Zheng, Y., Nozawa, Y., Gao, X., & Chang, H. H. (2012). *Multistage Adaptive Testing for a Large-Scale Classification Test: Design, Heuristic Assembly, and Comparison with Other Testing Modes. ACT Research Report Series, 2012 (6).* ACT, 500 ACT Drive, P.O. Box 168, Iowa City, IA.

BIOGRAPHICAL SKETCH

Halil Ibrahim Sari was born in Kutahya, Turkey. He received his B.A. in mathematics education from Abant Izzet Baysal University, Turkey. He later qualified for a scholarship to study abroad and in the fall of 2009, enrolled for graduate studies in the Department of Human Development and Organizational Studies in Education at the University of Florida. He received his M.A.E in research and evaluation methodology from the Department of Human Development and Organizational Studies in Education in August, 2013. He did his PhD in the same program.