DALLAS, ANDREW, Ph.D. The Effects of Routing and Scoring Within a Computer Adaptive Multi-Stage Framework. (2014)
Directed by Dr. Richard Luecht. 123 pp.

This dissertation examined the overall effects of routing and scoring within a computer adaptive multi-stage framework (ca-MST). Testing in a ca-MST environment has become extremely popular in the testing industry. Testing companies enjoy its efficiency benefits as compared to traditionally linear testing and its quality-control features over computer adaptive testing (CAT). Test takers enjoy being able to go back and change responses in review time before being assigned to the next module. Lord (1980) outlined a few salient characteristics that should be investigated before the implementation of multi-stage testing. Of these characteristics, decisions on routing mechanisms have received the least attention. This dissertation varied both item pool characteristics such as the location of information, and ca-MST configuration characteristics such as the ca-MST configuration design (e.g., 1-3, 1-2-3, 1-2-3-4). The results from this study hope to show that number correct scoring can serve as a capable surrogate for IRT calibrations at each step and that even if three-parameter scoring models are used at the end that the number correct method will not misroute as compared to traditional methods.

THE EFFECTS OF ROUTING AND SCORING WITHIN A COMPUTER ADAPTIVE MULTI-STAGE FRAMEWORK

by

Andrew Dallas

A Dissertation Submitted to
the Faculty of The Graduate School at
The University of North Carolina at Greensboro
in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Greensboro
2014

Approved by

Richard M. Luecht
Committee Chair

APPROVAL PAGE

This dissertation, written by Andrew Dallas, has been approved by the following committee of the Faculty of The Graduate School at The University of North Carolina at Greensboro.

Committee Chair Richard Luecht

Committee Members Terry A. Ackerman

Robert Henson

John Willse

_____

Date of Acceptance by Committee

_____

Date of Final Oral Examination

ACKNOWLEDGMENTS

I am grateful to my committee, family, friends, and colleagues who endured my questions, comments and frustrations. I received immense support throughout the course of my graduate school career from all four members of my committee. Dr. Richard Luecht assisted me both professionally and personally and through his mentorship has opened more doors for me than I could have imagined, even despite the constant ribbing that I endured for being a Bears fan. Dr. John Willse and I served on a grant together where I picked up enormous amounts of experience and ran more analyses than I could have ever possibly learned in the classroom, (even if many of them came up fruitless). Dr. Robert Henson served as my master's advisor and ensured that I was able to follow my own path in crafting the graduate school experience through the department. He also was able to see connections in the work that I was not able to see at the time. Finally I would like to thank Dr. Terry Ackerman whose door was never shut, even though I often would walk into his office simply to discuss events outside of academics. I could not have gotten nearly as far as I have without the tremendous support that I have received from my committee.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER I

INTRODUCTION

This chapter discusses the past and current state of computer adaptive multi-stage testing (ca-MST). Furthermore, this paper places ca-MST in the spectrum of other popular test delivery models. Next, some terminology of ca-MST is discussed, as well as, some of the key variables for constructing a study on ca-MST. Finally, a gap statement and relevant research questions are discussed.

An Overview of Some Relevant Test Delivery Models

Computer-based testing (CBT) is becoming widely accepted across multiple venues and platforms, with testing applications in military selection and placement, psychological assessment, educational assessment, language assessment and professional certification/licensure. There are many types of CBT delivery models, although most can be characterized by several key test delivery features: (a) whether the tests are pre-assembled or assembled in real-time (while or immediately prior to the examinee starting the test); (b) whether the test length is variable or fixed; and (c) the degree to which the test is made adaptive. Real-time test assembly, variable-length, and item adaptation features are most often associated with computerized adaptive testing (CAT). This section discusses the spectrum of CBT models available and places ca-MST along that spectrum.

CAT is typically advocated as a way to make a test more efficient (i.e., improve the stability of estimated proficiency scores over a non-adaptive test of the same length or maintain a prescribed level of score estimation stability or "standard errors of estimate" using a shorter test). A CAT works by essentially targeting the difficulty of each test item to the estimated proficiency of the examinee, updating the proficiency score estimate after each item is administered. The targeting algorithm provides the adaptive feature and occurs in real-time. In addition, a CAT can be variable length (e.g., stopping when a prescribed standard error of estimate is reached).

Pre-assembled, linear fixed tests (LFT) tend to be fixed in length and are non-adaptive. LFTs contain a fixed collection of items that every examinee assigned that form sees. They may be administered as a CBT or in paper-and-pencil format. What LFTs lose in efficiency by being non-adaptive, they tend make up for in terms of providing many important quality control (QC) and quality assurance (QA) advantages—including the capability to review and change (if necessary) the mix of items selected for each test form.

An alternative CBT delivery model that is gaining steady support in operational testing circles is called a computerized adaptive multi-stage test (ca-MST). Ca-MST blends the efficiency of CAT with the strong QC/QA capabilities of a LFT. The Uniform Certified Public Accountants Examination (American Institute for Certified Public Accountants Exam) was one of the first large-scale, high-stakes organizations to implement a ca-MST platform in 2004. More recently, the Graduate Record Examination (GRE®, Educational Testing Service) has replaced its CAT versions with ca-MST

versions. Other testing companies are also considering the transition to ca-MST because this test delivery models offers many practical benefits.

The simplest definition of ca-MST is that it is an adaptive test that employs pre-assembled, carefully structured test assembly units called *modules* that are comprised of multiple items and *panels* that are comprised of multiple modules. Each ca-MST panel is fully capable of self-adapting an examination form to each examinee—similar to a CAT—based on his or her performance over a series of two or more testing stages. The adaptivity is controlled by the *routes*, which guide examinees from module to module. Unlike CAT, which adapts at the item level, ca-MST adapts at the module level—but in a highly structured way. The design of the modules and panels achieves a certain amount of statistical efficiency via adaptive *routing* to a specific set of modules along the allowable routes in the panel. Furthermore, the pre-construction of every panel allows direct QC review and control over all of the content across any route (test form) in the panel (Luecht, 2000; Luecht & Nungester, 1998). The same level of strong QC capabilities is simply not feasible in CAT because of the real-time test assembly.

Despite its growing popularity, there are still some important technical research issues that need to be addressed for the full potential of ca-MST to be realized. These technical research issues include determining the appropriate number of stages and the number of difficulty levels per stage for various test purposes (i.e., optimal ca-MST configurations), measurement information targeting strategies, automated test assembly mechanisms and the impact of severe content balancing and item pool supply limitations, balancing information demands with item security and over- or under-exposure,

provisional scoring and routing mechanisms, and optimal pretest and calibration designs (e.g., Luecht, 2012; Willse, Ackerman, & Luecht, 2012; Zenisky, Hambleton, & Luecht, 2010). The present study focuses on one very pragmatic issue that has important operational ramifications for ca-MST: an investigation of provisional scoring and routing mechanisms. Additional details are provided further on in this chapter (see **Purpose and Rationale**).

Elements of ca-MST

This section intends to further define the elements that comprise a ca-MST and use a specific terminology introduced by Luecht and Nungester (1998). Figure 1.1 depicts multiple 1-3 panels at the left and multiple 1-2-3 panels at the right. Multiple items are assigned to pre-constructed *modules*. The length of each module can range from small (e.g., 5 to 10 items) to large blocks (e.g., 50 or more items). In addition, modules can be entirely or partially comprised of computerized performance-based tasks or simulations. In turn, the modules are assigned to difficulty-based slots in a self-contained test-administration unit called a *panel*. Each panel has a fixed number of stages in its design and each stage has a fixed number of difficulty levels (e.g., one to five levels of difficulty is common). For example, a two-stage 1-3 ca-MST configuration would have one module at Stage 1 and three modules (easy, medium, and difficult) at Stage 2. A 1-2-3 panel would have three stages: one level of module difficulty at Stage 1, two levels of module difficulty at Stage 2, and three levels of module difficulty at Stage 3. In addition, multiple panels are pre-assembled under the chosen configuration for security purposes—similar to have multiple test forms or versions of a test.

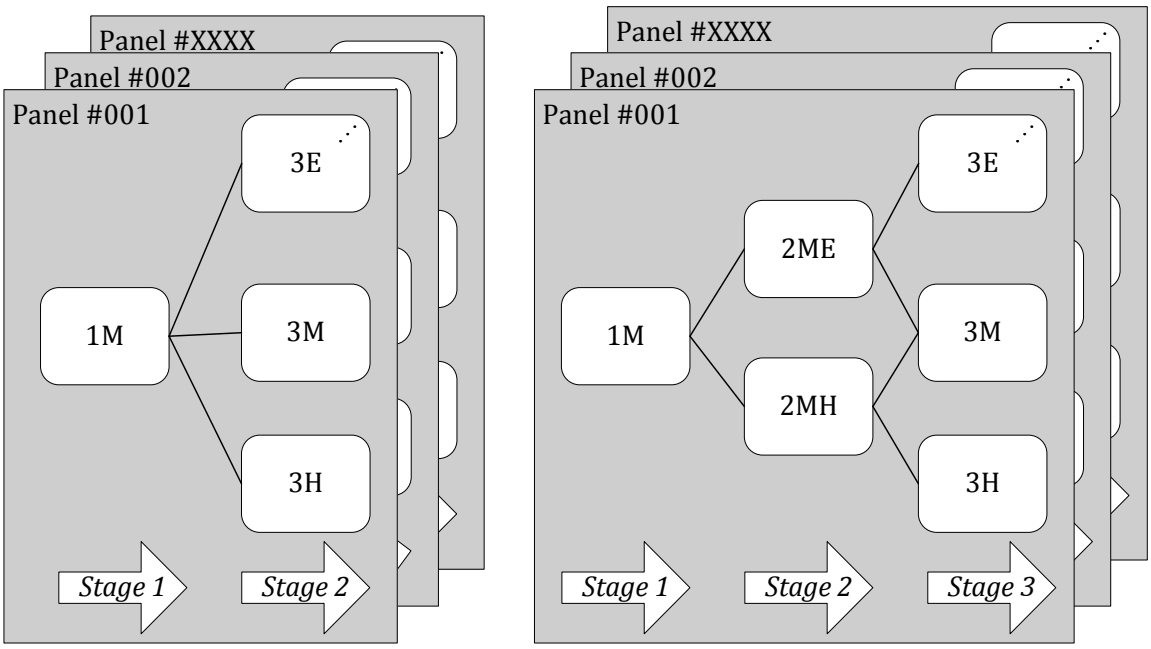

Figure 1.1. Two ca-MST Panels. (Left) A 2-Stage 1-3 Design; (Right) A 3-Stage 1-2-3 Design.

Each rounded-corner rectangle in Figure 1.1 is a *module* comprised of a fixed number of items. The letters and numbers respectively denote the difficulty level (E = easy, M = moderate, H = hard or difficult, ME = moderately easy, and MH = moderately hard). The solid lines dictate the allowable routes through the panel. Other *auxiliary* routes can be allowed as a matter of policy. Auxiliary routes allow examinees to jump in difficulty. In practices, all of the modules and all of the panels would be pre-assembled to meet fairly precise statistical targets, content requirements, and item-exposure restrictions. The parallel or near-parallel panels can then be randomly assigned to examinees, filtering out for repeat test takers any panels containing previously seen

modules. Once assigned, the ca-MST panel becomes a self-adapting test, justifying the "ca" in the acronym.

Scoring and routing is handled within each specific route. For example, a high-ability examinee might take the first (moderate) module in a 2-stage, 1-3 panel configuration design, and then be adaptively routed to the most difficult module on the second stage. Each examinee is routed through the stages to modules targeted in average difficulty to their proficiency as estimated from all modules administered up to that point in the test. Therefore, *routing* is synonymous with *adaptation* under ca-MST. Students that perform well on previous modules will be routed to more difficult modules; students who perform poorly on previous modules will tend be routed to the less difficult modules in that panel. *Primary routes* in each panel are the expected trajectories taken by most examinees that perform at a consistent proficiency level across all of the stages. Auxiliary routes allow for the possibility of recovery for an examinee near the routing cut points, or for examinees that for whatever reason exhibit somewhat more erratic performance from stage to stage. The extent to which an examinee can recover can be limited as a matter of policy by turning off particular routes. For example, to reduce the incentive for examinees taking to the test in order to memorize sets of items, examinees may be precluded from jumping from the moderately easy module at Stage 2 (2ME in the 1-2-3 panel configuration shown in Figure 1.1) to the most difficult module at Stage 3 (3H) as a matter of test security policy. The highly structured data design of each panel configuration makes these types of policy rules very simple to implement.

Benefits of ca-MST

This section describes some of the benefits of switching to a ca-MST including a balance between adaptivity and QC/QA processes found in LFT. This section also places ca-MST within the spectrum of assessment types.

There are several different ways to operationalize a CBT once the decision has been made to move in the direction. Each has its benefits and drawbacks. In terms of adaptivity on one end of the continuum is LFT, which includes both CBT and paper-and-pencil test forms. LFT is non- adaptive and every examinee assigned a particular form typically sees exactly the same test items. A real-time assembled LFT is called linear-on-the-fly testing (LOFT, see Folk & Smith, 2002). LOFT gives each candidate a unique set of items that fit content and statistical specifications. The test assembly process is carried out in real-time and usually combines a test assembly algorithm with item exposure control mechanisms—the latter of which add a random selection aspect to the test assembly process, making it impossible to exactly determine beforehand which items a particular examinee will see. The primary advantage to implementing a LOFT is its test security advantages insofar as preventing any examinees from being able to predict which test form or items they will see. However, LOFT is non-adaptive and the real-time assembly further precludes many QC/QA features available for LFTs (and ca-MSTs).

On the other end of the adaptive test assembly continuum lies CAT. As noted earlier, CAT is completely adaptive at the supposedly lowest unit of test delivery—the *item* (although, some have even advocated for self-adaptive items—see Luecht, 2009, 2012). An initial item or small set of items is administered by some random or other

mechanism to begin the CAT. Once the CAT sequence begins, each subsequent item is specifically selected from the item pool to maximize the measurement information (reciprocal of the error variance of estimation) for that examinee's current proficiency estimate. The CAT heuristic can incorporate content balancing constraints or other mechanisms, and provides a variety of stopping options, such as stopping when a desired level of score stability or decision accuracy is reached (Jodoin, Zenisky, & Hambleton, 2006; van der Linden, 2005; Veerkamp & Berger, 1997). However, because CAT performs all of the test assembly operations in real-time, it provides test developers the least amount of control over the actual quality of test forms. At best, simulations can be carried out to mimic the adaptive process and/or post-test administration reviews of test forms can be conducted by subject matter experts. In addition, because of the tendency for the CAT heuristic to always choose the statistically "best" items, item-exposure control mechanisms or item-exposure constraints are usually needed to offset the continued reuse of the same high-demand items for large numbers of examinees. These item-exposure controls partially solve an inherent item security risk problem associated with CAT, but then may choose statistically suboptimal items (Revuelta & Ponsoda, 1998; Stocking & Lewis, 2002; van der Linden & Veldkamp, 2004). Other exposure control methods such as the Weiss (1973) match-to-*b* or Chang, Qian, and Ying (2001) which combined a-stratification with Weiss's b-blocking have shown promise of reducing item exposure while maintaining efficiency. Chang et al. (2001) acknowledge that further work needs to be done in exploring what types of item bank would be acceptable for match to b stratification, also, the number of strata to be used. While these

are all legitimate limitations under CAT, it still remains the single-best way to optimize the statistical precision of the score estimates—at least as measured by the error variance of estimate or amount of statistical measurement information provided.

Ca-MST flexibly lies somewhere in the middle of the adaptive continuum, ideally combining the best features of each extreme. In fact, ca-MST configurations and measurement information targets can be designed for almost any application or test purpose. For example, ca-MST designs with many stages and levels of difficulty (e.g., a 5-stage 1-2-3-4-5 panel design) would move closer along the continuum toward item-level CAT. A simple 2-stage 1-2 panel design would be less adaptive and closer to a LFT.

The apparent benefits of ca-MST can be realized because of two important design features: (a) high-integrity, hierarchically oriented data structures and (b) pre-assembly of all of the test units (modules and panels). High-integrity data structures provide for seamless control of how the modules are constructed, referenced and controlled from initial test assembly through processing of the response data. Pre-assembly of the modules and panels also affords any desired level of "content review" as well as automated QC/QA procedures to determine exactly which items appear in every finalized module and panel, how items are shared across modules (if at all), how modules are shared across panels, how the routing within a panel occurs, and even allows test administration features such as timing controls, item sequencing within modules and item review to be precisely controlled.

From a strict psychometric perspective of statistical efficiency, these two design features may seem to be somewhat trivial, however, for large-scale, operational CBT they most assuredly are very important by challenges from subject matter experts. For example, the same typical subject-matter expert (SME) review processes that are often used for LFT can be used with ca-MST. The panels can be completely content-balanced and checked for relevant statistical qualities before any panel is released for operational use. Also, item exposure can be completely controlled with risks known *a priori* (Dallas, Wang, Furter, & Luecht, 2012; Luecht, 2003; Luecht & Burgin, 2003). Unlike CAT, which has the potential to overexpose particular "high demand" items (e.g., items with high discrimination parameter estimates when the two- or three-parameter IRT models are used), ca-MST modules and panels are pre-constructed to match prescribed measurement information targets and simultaneously control item exposure within a particular population of examinees. The routing through the ca-MST panels also precludes certain types of behaviors by examinees who attempt to "game the system" (e.g., performing poorly early on in hopes of getting easier test items and then streaking to the finish with near perfect performance). Finally, ca-MST is adaptive and therefore results in more precise (stable) score estimates than LFT. While perhaps not as precise as item-level CAT, a practical level of precision can be achieved easily by using sensible ca-MST test assembly measurement information targets and sound item writing/inventory control to ensure that the supplies fulfill the demands.

Purpose and Rationale

This section outlines the state of research in the field of ca-MST and discusses the current gap that this study attempts to fill. In addition, this section outlines the rationale for examining number correct scoring in practice.

Most of the foundational research on ca-MST to date has focused on basic ca-MST designs features and efficiency comparisons with both LFT and CAT (e.g., Jodoin et al., 2006; Luecht, 2000; Luecht, Hadadi, & Nungester, 1996; Luecht & Nungester, 1998; Reese, Schnipke, & Luebke, 1999; Schnipke & Reese, 1999; Xing, 2001; Xing & Hambleton, 2004). There are relevant findings from that research. For example, as the length of the test gets longer precision increases (Luecht et al., 1996; Luecht & Nungester, 1998; Jodoin et al., 2006; Wang, Fluegge, & Luecht, 2012). In fact, Luecht et al. (1996) demonstrated that three-stage ca-MST was practically as efficient as an item-level CAT when "typical" test lengths of 30 or more items were used. Similarly, Jodoin et al. (2006) found that there are upper limits on efficiency gains from increasing the test length beyond 40 items. More specifically, he show that a 40-item ca-MST performed almost as well as 60 item tests in terms of measurement precision (standard errors and score accuracy). Other research studies have also compared the relative efficiency of ca-MST versus CAT or LFT (Jodoin et al., 2006; Kim & Plake, 1993; Luecht, 2006; Reese et al., 1999; Schnipke & Reese, 1999; Xing, 2001; Xing & Hambleton, 2004). These studies typically employ large-scale computer simulation to both compare the test delivery models, statistical targeting strategies for test assembly, and test lengths, as well

as to highlight the conditions under which particular LFT or ca-MST designs or CBT delivery models produce inaccurate scores or decisions.

Zenisky et al. (2010) indicated two important areas of needed future research: (a) better understanding the inherent interactions between the ca-MST panel test assembly demands and the available distribution of item information and content in the item pool (supply), and (b) understanding the impact of different test routing and scoring strategies on examinees' scores. Recently some researchers have begun to take on these two issues as part of a more elaborate research agenda. For example, Luecht (2012), Willse et al. (2012), Wang et al. (2012), and Dallas et al. (2012) manipulated large numbers of ca-MST design configurations ranging from two to five stages, included assembling and administering variable numbers of panel replications from finite sized item banks and for multiple test lengths, and introducing various item pool characteristics (i.e., item discrimination and difficulty parameter distributions plausibly modeled after existing item pools to specifically look at the interactions between ca-MST design demands and item pool inventory supply issues). These studies found that the item pool characteristics can directly or indirectly affect how well the ca-MST assessment system operates to achieve efficient test administration and scoring. Most notably these researchers discovered some important item pool design conditions under which ca-MST might start to operate at a suboptimal level—from a psychometric perspective. For example, if the test information function for the item pool is highly leptokurtic at some point of the proficiency scale, a more adaptive 1-3-5 panel design might not be the best design since supply of measurement information will be insufficient to support five distinct levels of

difficulty at the third stage. This same supply-and-demand problem occurs when CAT is implemented using an item bank that was originally developed to only support [non-adaptive] LFTs. But the supply and demand problem in a non-adaptive LFT might not be as clear due to the real-time test assembly and exposure controls used. Because the test assembly statistical measurement information function targets used in ca-MST are [usually] explicit (Luecht, 2000, 2012), deficits in the supply, given these types of demand-supply interactions are relatively straightforward to investigate.

This study addresses both of the topics specified by Zenisky et al. (2010). This type of research has a practical focus, ideally providing testing companies with empirical evidence—or at least suggesting sound investigative methods—to examine their own item pools as *supplies* relative to their test design and assembly *demands*. Based on the results, they can either invest in the necessary item writing activities to improve the supply side over time, or modify their demands to match what is practically achievable.

As noted above, there is still a serious and apparent need for research to further evaluate the effect of routing decisions on overall accuracy and precision of measurement. Although it is possible to exclusively rely on IRT for building the ca-MST modules and panels, scoring the examinees, and routing them using a maximum-information module-selection criterion similar to CAT, it is also possible to simplify the routing and module selection by using IRT to assemble the modules and panels, but then convert to number-correct (NC) scoring and NC routing tables for operational, real-time movement within a panel (Luecht & Nungester, 1998; Luecht & Burgin, 2003; Luecht, Brumfield, & Breithaupt, 2006). IRT-based routing and scoring mimics CAT by

selecting modules to minimize scoring estimation errors and actually estimating maximum likelihood or Bayesian proficiency scores in real-time. When large numbers of examinees are testing at one time, IRT routing and scoring can introduce severe processing loads on internet servers, or introduce security issues if the entire process is downloaded to a local workstation or local network file server (Luecht, Brumfield, et al., 2006; Zenisky et al., 2010). In contrast, NC simplifies operational scoring and routing demands and can substantially reduce various data transmission and processing loads for an operational ca-MST system. However, the utility (costs and benefits) of these scoring and routing alternatives have not been systematically explored from a psychometric perspective. The purpose of this study is to explore whether the overall efficiency of NC scoring provides enough benefit to outweigh the cost in scoring accuracy.

This research study examines both the impact of NC versus IRT-based routing and scoring methods on the accuracy of scores and the various other item pool and ca-MST design factors and conditions that can interact to also affect score-accuracy outcomes. This study does use simulated data, but the basic distributions of item characteristics are modeled after a large-scale, operational item pool. This study includes manipulations of these characteristics that are on the edges of what is commonplace in testing; however, these liberties are taken in order to explore the potential impact of having rather extreme demand/supply deficit conditions. In that regard, this study not only evaluates the technical viability of NC routing for ca-MST designs, but also highlights conditions or scenarios that could prove problematic in practice.

Study Variables, Factors, and Issues

The study variables can be broken up into two categories: (a) ca-MST design features, and (b) item pool conditions. The ca-MST components are variables such as the module length, the ca-MST design configuration (i.e.: number of stages, levels of difficulty per stage, etc.) and the method of routing examinees. Item pool conditions include the joint distributions of item discrimination and difficulty in the item pool. Ultimately, it is the interaction of these two categories that are of interest. Computer-based simulations are used in this study to specifically manipulate those interactions and evaluate (a) comparatively which method of routing and scoring works best under which conditions and (b) under which (if any) studied conditions either or both methods of routing and scoring produce marginal of substandard results.

Test length is manipulated via the module length. Few operational tests (at least high-quality, high-stakes exams) are shorter than 30 items. However, because this is an adaptive test and a simulation study, it makes sense to investigate how small the test can become before recovery of the generated "true scores" is jeopardized. The longest tests considered in this study somewhat mimic some of the large-scale professional certification exams where hundreds of items are given to examinees to ensure that the candidates' certification will not result in a danger to the public. In sum, only two module lengths are considered (10 or 20 items), however, when fully crossed with the three ca-MST panel configurations, there are six total combinations of ca-MST design features.

Ca-MST designs are plentiful and can take on almost any desired characteristics by manipulating the number of stages and levels of difficulty within stages. Obviously, the design demands should be consistent with the available supply of items (measurement information and content) in the pool. In this study, some rather common ca-MST panel designs in the research are considered, as well as, a few more exotic designs. Previous studies (Jodoin et al., 2006; Luecht, Brumfield, et al., 2006; Luecht & Nungester, 1998; Reese et al., 1999; Xing & Hambleton, 2004) considered panel two- and three-stage panel configurations such as 1-3, 1-3-3, and 1-3-5. Other more exotic designs have also been used in practice such as the 6-stage 1-5-5-5-5-5 used by Crotts, Sireci, and Zenisky (2012). The present study considers three specific ca-MST designs (see Chapter III).

In addition, routing decisions can influence both precision of measuring the intended proficiency and item exposure in a ca-MST assessment system. This study examines four different routing and scoring methods in terms of two interacting selection and scoring function factors: (a) the module-selection criterion, and (b) the type of scoring used for the actual routing. Table 1.1 illustrates the interacting selection and scoring function factors.

There are two ways to select modules: (a) using a maximum-information module selection criterion, computed at some provisional proficiency score, to minimize the error of estimate, or (b) routing based on fixed cut score points to achieve desired item (module) exposure levels based on an assumed distribution of examinees. Luecht and Burgin (2003) also discussed a hybrid method of designing information targets to simultaneously achieve both goals; however, their approach will not be considered.

There are also two scoring methods that can be applied with either one of these two module-selection methods: (i) IRT scoring using either maximum likelihood or Bayes estimates of the proficiency scores, θ; or (ii) number correct (NC) score routing using a routing table. The four routing and scoring methods, summarized in Table 1.1, investigated in this study represent the combinations of module-selection and scoring methods.

Table 1.1

Table of Routing and Scoring Conditions

		Module Selection Criteria	
		Maximum Information Selection	*Exposure Control Selection*
Type of Scoring Function	*IRT Scoring*	Maximum Information Selection with IRT Scoring	Exposure Control Selection with IRT Scoring
	Number Correct Scoring	Maximum Information Selection with Number Correct Scoring	Exposure Control Selection with Number Correct Scoring

This study also examines eighteen different item pool characteristic conditions as function of the distributions of item difficulty and item discrimination. In fact, perhaps the only limiting factor in this study is that all of the examinee proficiency scores are sampled from the same unit normal population, $\theta \sim N(\mu = 0, \sigma^2 = 1)$. The potential interaction between examinee density, test purpose and measurement information (precision) density, however, is indirectly addressed via the various item pool characteristics. More specifically, this study includes nine difficulty distributions

specified as a function of three mean item difficulties, μ(*b*), and three levels of standard deviation, σ(*b*). Those nine item difficulty distribution conditions are also fully crossed with two levels of average item discrimination. Controlling the item pool difficulty and discrimination distributions allows for interactive comparisons of the routing and scoring methods under different inventory supply contexts.

Multiple data set replications are generated under each of the 270 design conditions: 15 ca-MST design conditions by 18 item pool characteristic conditions. (Note that the multiple replications of the data sets under each condition will provide empirical sampling distributions of various aggregated outcome statistics reflecting potential bias and the variation of estimation errors.) Large-sample data sets are also used to ensure sufficient distributions of examinees across the proficiency score scale. Each of the four routing and scoring mechanisms are applied to each data set using ca-MST simulation software (i.e., the data sets are held constant so that the simulated examinees act as their own controls across the four routing and scoring methods

Research Questions

1. How do the four routing and scoring mechanisms impact the recovery of simulated proficiency scores under a variety of ca-MST conditions?
2. Under what conditions do various ca-MST design configurations work best from a psychometric scoring-accuracy perspective with respect to the four routing and scoring mechanisms as well as the various item pool characteristics?

3. How do various item pool characteristics impact the quality of ca-MST test forms from a psychometric perspective?

All research questions will be addressed using a simulation approach that will include a fully crossed design with 270 distinct conditions, 10 replications of each set of conditions, and using the same examinees under each of the four methods of routing and scoring. It may seem important to re-emphasize this latter point about holding the examinees constant for each method of routing and scoring. That capability helps to directly control for sampling errors in the head-to-head methodological comparisons.

Summary and Organization

With the push to use technology more and more in classrooms, assessment designers and test developers need to react to the ever-changing world in education and other areas of testing. Ca-MST presents one very appealing way to incorporate cutting-edge technological enhancements in CBT with a sound adaptive test delivery strategy, while still retaining full quality control (QC) capabilities regarding items, test forms and response data. This latter QC issue is simply not well addressed with real-time test assembly algorithms such as those employed with CAT or LOFT. The flexibility of ca-MST configurations, the pre-assembly of the modules and panels, and the use of high-integrity data structures underneath the apparent surface; make ca-MST a very robust way to implement an effective adaptive testing program. Testing survey research has also demonstrated that examinees like the capability to review and possibly change their answers within a module. Other research has shown that some of the more flexible three-

stage ca-MST designs come very close to achieving the statistical efficiency of item-level CAT, especially for tests of realistic length (e.g., 30 or more items).

This study will investigate precisely the issues that Zenisky et al. (2010) flagged as areas of future research. Most notably a deeper understanding of the interactions between the ca-MST assembly and the available information in the item pool as controlled by the study pool conditions (e.g., the effects of the item pool on estimation accuracy, Dallas et al., 2012; Wang et al., 2012). In addition, this study is an in-depth exploration of the impact of alternative methods of test routing and scoring strategies and their influence on examinee scores.

This study will contain five chapters in total. Chapter I provided a rationale for this study. This chapter describes the topic, existing research, and gap(s) this study helps to address. Chapter II is a description of the literature on the topic of ca-MST and other areas. The literature review elaborates on the issues discussed in Chapter I, as well as, discuss all research related to ca-MST's including efficiency, estimation, routing, and panel construction. Chapter III contains the methods for this study including descriptions of the conditions, method of the simulation, relevant information about each data set, and relevant analyses. Chapter IV describes the results including figures/plots and other relevant results. Finally, Chapter V discusses the results and the implications the results have for ca-MST research.

CHAPTER II

REVIEW OF THE LITERATURE

This chapter reviews the literature related to ca-MST's including a review of IRT. The chapter discusses the literature relating to the six topics that Lord (1980, as cited in Zenisky et al., 2010) argued were the key technical decisions needed to develop a MST, compares ca-MSTs to other test delivery models, describes item pooling for both a ca-MST and other test delivery models, and a review of Item Response Theory including a discussion of the potential models for this paper. These topics inform the research questions by illustrating both the focus of research over the past two decades in ca-MST but also the gap that currently exists in the field.

Lord's Six Areas of Research

Ca-MSTs were first theorized by Lord (1980) when Lord outlined six key factors previously discussed for setting up and implementing a multi-stage test. Lord's work revolved around setting up a two-stage test and references Cronbach and Gleser (1965, Chapter 6) using a decision theory approach. Ultimately, Cronbach and Gleser sought to deal with a situation such as a certification examination where a single cut existed and a decision needed to be made. To them, the 2nd stage was only necessary to clear up borderline decisions. Lord's work dealt more broadly with the rationale of performing good measurement as opposed to clarification of a classification. Lord mentioned several issues test developers would need before developing a ca-MST. Lord, however, did not

specify how to go about setting up a multi-stage test nor lay out a principled approach to MST.

The six topics that Lord (1980, as cited in Zenisky et al., 2010) outlined were:

- Total number of items in the test
- Number of items in the initial and each n-stage module
- Difficulty of the initial module
- Number (and difficulty) of alternative modules in each stage
- Cut points for routing examinees to modules
- Methods for scoring stages and each nth stage test.

This study directly examines the last two bullets. This section will discuss the literature existing in the field on ca-MST paying special attention to the decisions made pertaining to these topics.

A closely related area to ca-MST is the testlet models discussed by Wainer and Kiely (1987). Wainer and Kiely defined "testlets" as a group of items that relate to a single content area and are analyzed as a unit. While these are similar to item modules, testlets are typically considered related within a module above and beyond random chance. An example would be a set of items that are related to the same reading comprehension passage or questions about a graph/diagram (Hendrickson, 2007). Ca-MST's (modules) can be testlets however do not make the strong assumption that the items are based strongly on a central idea or goal but rather are more general.

Luecht and Nungester (1998) described a framework for a large scale computerized testing framework that was called Computer Adaptive Sequential Testing

(CAST). In addition, this research set up the language referred to in ca-MST, panels as preconstructed units and modules as sets of items assigned to a particular stage. Luecht and Nungester argued that the best way to set up this design was to "consistently match multiple statistical targets" (p. 230). The best way to do this is via an ATA engine. In using an ATA engine, the panels and modules could be built specifically to incorporate both statistical targets, such as information or low standard error, but content demands as well. This will be discussed further in the **Strategies for Item Pooling** section. Luecht and Nungester (1998) had 180 total items (90 total items per module in the 2-stage, 60 total items per module in the 3-stage) in their study. Information or the method of conditionally controlling precision across the score scale were created using Test Information Targets (TIF). While TIF targets were initially set the process of selecting items had a difficult time finding items to fit the hardest modules as the item pool had not set out with this goal in mind. This study investigated three different ca-MST designs: 1-3, 1-3-3, and 1-3-5.

Reese et al. (1999) compared a 2-stage testlet design to their current practice of a paper pencil design. Testlet designs differentiate themselves from a ca-MST because typically testlets are centered on a central reading passage and thus have a high level of association between their items. While all testlets can be considered modules, not all modules could be considered a testlet. This study involved a 1-3 design that had 10 items in the first stage and 15 items in the second stage. The testlets were developed so that the first stage had lower discrimination values and an approximate average b of $N \sim (\mu = -0.50, \sigma = 0.80)$. The second stage the difficulty of the low module was at $N \sim (\mu = -1.00,$

$\sigma = 0.80$), the medium module was $N \sim (\mu = 0.00, \sigma = 0.80)$, and the high difficulty module was $N \sim (\mu = 1.00, \sigma = 0.80)$.

Xing and Hambleton (2004) on the other hand used 35 items total for an information systems examination. A 2-stage ca-MST design was used which used different length first and second stage module lengths as the routing test was slightly longer than the 2nd stage module (15 items - 20 items). Xing and Hambleton (2004) noted that the total number of items (35) was on the "small side" for credentialing exams; however it is not entirely uncommon within the information technology field for credentialing exams.

Several studies have gone into judging the efficiency of multiple test lengths. For example, Jodoin et al. (2006) compared two and three stage ca-MSTs to a LFT and several operational forms in terms of decision consistency. Four operational forms were considered the "real" condition. Then, from the item pool three new 60-item LFT's were created. Next, 60-item 3-stage (1-3-3) ca-MSTs were created and 40 item 2-stage ca-MSTs were created. In both the 3-stage ca-MST and the 2-stage ca-MST, the modules in each stage contained 20 items each. The ca-MSTs were created in two ways. Method 1 was built so that each module contained a third of the information of the 60-item LFT at each stage. Method 2 was built so that the first module was reduced to a quarter of the total information of the LFT, and the latter two modules were three eighths each.

Each of the 3-stage ca-MST outperformed the LFT's while the 2-stage results were comparable or a little worse than the full 60 item LFT. As the pass rate was increased from 30% to 50%, LFT seemed to close the gap between it and the three-stage

ca-MST. In addition, Method Two seemed to do better in the 3-stage design; however, Method Two seemed to do worse overall in the 2-stage design. The current operational forms did comparably well to the 3-stage design and overall did the best except when the pass rate was raised to 50%. In the 30% and 40% conditions it was only marginally better.

The results from this study can help clear up two issues. First, the 2-stage 40-item test performed worse than the 60 item tests; however, not a great deal worse. Considering the time savings and money savings that can be achieved with a far shorter test, these results are promising. In addition, when there are more than 2 stages, shorter tests are satisfactory as long as the adaptivity of the 3-stage form is allowed to work. This finding in a way disagrees with Lord's premise that short routing stages can end up causing problems. If there are enough stages to adequately recover it looks like the routing test length can be short and the latter stages a little bit more informative.

Other studies such as Crotts et al. (2012) have supported the study done by Jodoin et al. (2006) and showed that a reduction from 40 items to 35 items has a negligible effect on the six-stage Massachusetts Adult Proficiency Test for Reading. Overall, the reduction in reliability (Cronbach's α) was minimal dropping from roughly .93 to .90. In addition, classification consistency was comparable to the 40-item test. This resulted in a savings of about fifteen percent in test-taking time, which is an important feature when paying for a seat for the length of the exam.

These studies primarily alter the first four bullet points in Lord (1980) considerations for building a MST. Most studies played with the number of items in the

test and the number of items in each n-stage module. While not directly, the studies usually did have the initial module in the middle of the item pool distribution; however, this isn't necessary but makes the most sense. Each study used a conservative design as well, the 1-3 (Luecht & Nungester, 1998; Schnipke & Reese, 1999) and 1-3-3 (Luecht & Nungester, 1998); however, more recent studies such as Crotts et al. (2012) used a 6 stage design. This study alters some of these characteristics in a way consistent with the literature. This study considers item sizes that range from small to very large and considers multiple designs both well researched and new and innovative.

Comparison of Test Delivery Models

This section discusses the typical test delivery models versus newer adaptive models. The three delivery models discussed in this section are: LFT, ca-MST, and CAT. These methods are tested against each other in how well they are able to recover ability estimates (Reese et al., 1999; Patsula & Hambleton, 1999; Xing & Hambleton, 2004).

Reese et al. (1999) took the idea for a testlet based design and implemented a small study in practice that compared their LFT and testlet-based results. In a sense this research's primary focus was whether a ca-MST could handle complex content constraints without compromising efficiency that is gained over LFT and whether ca-MST increased the precision of the examination. This research found that the ca-MST was more precise than the LFT for most of the scale; however, at the extremes ($-2.5 > \theta > 2.5$) of the scale the LFT contained less bias than the two-stage ca-MST. Part of the lack of precision in the extremes was that the panel test information functions were not met to an acceptable degree for some of the panels in the design. One solution for this issue is

to lower the upper bound to spread information more evenly across panels. In addition, the authors argue that perhaps adding a 3rd stage one with a very low and very high module would address the issues with the lack of precision at the extremes. Overall, the RMSE for the two-stage ca-MST far exceeded both the 51-item LFT and the 25-item LFT.

Patsula and Hambleton (1999) evaluated various multistage testing models and compared these with LFT's and CATs. The conditions this study manipulated were the number of stages, the number of modules per stage and the number of items per module. This simulation resulted in the finding that ca-MST's end up with similar proficiency estimation when the number of stages in a ca-MST is increased from 2 to 3.

Xing and Hambleton (2004) examined fixed LFT's versus a two stage ca-MST and a CAT. The study manipulated two factors: the size of the item bank and the quality of the items in the item bank. The bank contained 240 items and the average discrimination for the 240 items was 1.0. The poor quality item bank and the high quality item bank had an average discrimination value of 0.60 and 1.40 respectively. While an average discrimination value of 1.40 may seem unrealistic, the authors note that these values were an attempt to examine the impact of a set of items with unusually high levels of item discrimination. In addition, the authors also manipulated the size of the item bank by doubling the amount of items in the bank.

Wainer and Kiely (1987) discussed earlier examined how a testlet based approach functioned and had two main findings. First, as the quality of items increased statistics such as the decision accuracy and decision consistency also increased. Better items

meant better precision which meant better estimation of ability. Secondly, the size of the pool was also important, in fact, more important than having highly discriminating items. The decision accuracies for the 480 size item bank and the original item discrimination, average discrimination = 1.0, was equivalent or better than the 240 item bank example with average discriminations of 1.4.

These studies indicated the efficiency and accuracy that ca-MSTs have over LFTs. While ca-MSTs may not be as efficient as CAT for shorter test lengths, these test lengths need to be considered to examine the strength adaptivity of the ca-MST even when the number of items is small. In addition, this section pointed out the effect of the item bank (more on this in the item banking section) on the overall success of the ca-MST. This study includes two different item banking qualities and 240 total different item banks.

Item Response Theory

Presently, most testing practices and assessment rely on Item Response Theory (IRT) to scale their assessments. The last thirty years, IRT has been the core of testing programs and represents an upgrade over the previous scaling methods such as Classical Test Theory (CTT). IRT is so useful because not only can all items be scaled to the same scale as people, but no longer is there a direct need to form to form equate as the parameters are considered to be invariant between samples.

In order to estimate an IRT model three major assumptions must be met. The first assumption is that a primary trait or dominant dimension accounts for the performance on any item or set of items. While multidimensional IRT (MIRT) models have been

developed, most IRT programs, including essentially all operational testing programs, assume that only one ability or trait is being measured at a time. Hambleton and Swaminathan (1985) makes the argument that unidimensionality can never strictly be met as there are other factors such as personality and test-taking factors that impact test performance to some extent.

The next assumption of IRT is Local Independence (LI). This assumption states that an examinee's responses to different items in a test are statistically independent. For this to be held true, examinee's performance on one item must not affect their responses either for the better or for the worse, to any other item on the test. It is important to denote that pairs of items should be correlated without first removing the effect of ability from the correlation. High ability examinees when faced with a pair of items should get both items right. Low ability examinees when faced with a pair of items should have consistent results. LI is only violated when associations between items are correlated at a fixed ability level.

The last assumption is that the model is correctly specified. There are three primary unidimensional IRT models that are used in common practice. There are three parameters that are fixed or let free in order to create the models. The three variables are parameters that represent difficulty, commonly referred to as the b-parameter; discrimination, commonly referred to as the a-parameter; and a pseudo-guessing parameter, commonly referred to as the c-parameter.

The first model is commonly referred to as the Rasch (Rasch, 1960) or 1PL model. The model estimates two parameters, the b parameter andΘ, a measure of ability.

The Rasch model is preferred by test developers because (a) all items are equally weighted, and (b) the model is considered to be statistically sufficient. In other words, the raw score is all that is needed to decompose the ability metric, θ. The Rasch model is expressed mathematically as

$$P(X_i = 1|\theta) = (\frac{\exp\{D(\theta - b_i)\}}{1 + \exp\{D(\theta - b_i)\}} \quad \text{(Eq. 1)}$$

DeMars (2010) makes the argument that all of these models are nested within each other, as a 2PL model is simply a 1PL without fixed a parameters and the difference between the 2PL and 3PL is simply removing the constraint of a 0 c-parameter (pseudo-guessing). The 3 PL formula is expressed as follows

$$P(X_i=1|\theta)=c_i+(1-c_i)(\frac{\exp\{(D*a_i)(\theta-b_i)\}}{1+\exp\{(D*a_i)(\theta-b_i)\}}, \quad \text{(Eq. 2)}$$

where D is a scaling factor either 1.0 for the logistic model or 1.7 for the normal-ogive model.

One of the bigger disagreements in model selection is whether or not the c-parameter should be included (DeAyala, 2009; DeMars, 2010). The pseudo-guessing parameter assumes that low ability examinees have a higher chance of getting an item correct than the 2PL model suggests. The parameter's estimation; however, is completely independent of this notion.

Typically, the more parameters that are included in a model the better the model will fit the data. To this end, some suggest that the 3PL, therefore should always be the

model of choice. Most often, the pseudo-guessing parameter does soak up random variation that the 2PL would not be able to capture. While it isn't necessarily related to guessing improving absolute model fit is never a bad thing. That said, what this premise ignores is that there is often an empirical association found between the discrimination parameter and the pseudo-guessing parameter estimates. This association between a and c may result in the 3PL being "over-parameterized" (Holland, 1990).

De Ayala makes the point that the c-parameter may actually pick up some person characteristics that it does not intend to such as "test-wiseness" or "risk-taking" tendencies. Therefore, the pseudo-guessing parameter isn't solely an item parameter, but also a person parameter or at the very least an interaction between item and people characteristics. Therefore, De Ayala (2009) makes a reasoned argument for why the 2PL might be a better model to select. De Ayala argues that the presence of the pseudo guessing parameter assumes the propensity to guess is constant across θ. This assumption may or may not be reasonable and there's not a really direct way to test this assumption. In addition, non-zero pseudo guessing parameters lower the estimate of θ (Wainer & Thissen, 1987), as well as, lower the amount of information that an item provides (De Ayala, 2009). Although, the non-zero guessing parameter does lower information, is it because information is truly lower or is information inflated in the 2PL because it ignores the concept that lower ability examinees have a non-zero probability of getting an item correct.

Despite these limitations, the 3PL is one of the most used models in psychometrics today. Many testing programs select prefer the lower complexity of the Rasch model and excellent available software, WINSTEPS, for operational work.

One of the advantages that IRT provides is the ability to predict examinee performance on items that they examinee has not yet seen. The manner in which this is accomplished is through estimation of an ability parameter, θ. There are multiple methods to accomplish the estimation of ability. This section of the paper will describe the two most popular methods of estimating θ, Maximum Likelihood Estimation (MLE) and Expectation A Priori (EAP) estimation. MLE and EAP are equally plausible methods for calculating θ. This section will describe how MLE and EAP estimation work briefly and then compare and contrast the differences between the two methods of estimation.

No matter which method you use to estimate θ the first step is calculating the likelihood function. The simplest method for estimating it would just be to determine the probability of the response vector given some known item parameters (ξ) and an estimate of ability (θ). The likelihood of an examinee's observed response vector can be calculated as

$$L(x_i|\Theta_i, \xi) = \prod_{j=1}^{L} p_j^{x_{ij}}(1 - p_j)^{(1-x_{ij})} \quad \text{(Eq. 3)}$$

(De Ayala, 2009, Hambleton & Swaminathan, 1985). This method will return the theta at which the likelihood is maximized. This method however is a crude approach to computing the likelihood. A more sophisticated likelihood is calculated instead. One

such method of computing the likelihood is MLE. This likelihood of a function has a few neat characteristics which makes this a much better procedure. First, near maxima of a distribution, the first derivative goes near 0 making it very small. In addition, at this same point the second derivative is large comparatively to the first derivative (DeMars, 2010) as a result the ratio of the first to the second derivative will be very small. The signs will be the same if the estimate of ability is too high, and will be opposite if the estimate is too low. There are multiple ways to start this procedure by giving it a starting value of a transformation of the number correct score or simply just starting at 0 (DeMars, 2010). Once it is started it will converge to some criterion set such as .01 (DeMars, 2010) and ability is estimated.

Bayesian estimation does nothing different computationally than MLE, the first and the second derivatives are still calculated and the convergence is achieved at some set criterion. EAP estimates are different; however, because they multiply a prior distribution to the likelihood distribution and attempt to find the maximum of the product of the two called the *posterior* distribution. The strength of the prior distribution can pull the likelihood function based on (a) how good the prior is, and (b) the strength of the prior. However, even when an inappropriate prior is used EAP and MLE tend to work equally well (Chen, Hou, & Dodd, 1998; Chen, Hou, Fitzpatrick, & Dodd, 1997).

Gorin, Dodd, Fitzpatrick, and Shieh (2005) performed a study using polytomous items with three types of priors, skewed positively, skewed negatively and normal prior distributions in a CAT procedure with three types of CAT item pools, middle peaked, peaked in the low end, peaked at the high end. This research found that MLE and EAP

estimates perform equally well under optimal item pool situations; however, when item difficulty does not match the items given to the examinees, EAP estimation may be the more appropriate method of estimation.

Generally speaking, MLE estimates tend to be biased outwards in that as theta tends to reach maximum scores, a max is usually just placed on these individuals. EAP estimates usually have a normal prior place on them so the tails might be biased a little inwards. Bock (1985) showed that in an adaptive environment and confirmed the inward bias. He noted that if the reliability of the assessment is sufficient; however, that the bias of the EAP is minor for most of the population (within +- 2 SD). He further elaborated stating that personnel decisions that were based on EAP estimates were equitable given that people competing for the same positions are estimated with the same prior and the same error criterion. In this study, EAP estimates will be used as they provide similar estimates under optimal conditions, and better estimates under sub optimal conditions.

IRT allows for conditional information over the entirety of the score scale. While CTT has measures of score precision (such as Cronbach's α) IRT allows information targets to be precise and precision can be known a priori when building ca-MST modules and panels. The NC routing, however, takes a step back, perhaps when routing in a ca-MST all one needs is the raw number correct in order to route examinees to the best panel for them. If this is true, then IRT is only needed to scale the assessments and not to route the examinees through the ca-MST. This study investigates whether or not the routing aspect of the ca-MST or the adaptive element can simply be done using a NC method.

Strategies for Item Pooling

There are three important topics discussed in this section. First, will be a presentation of a couple of competing theories on constructing and maintaining item pools. The next topic will be a discussion of Automated Test Assembly (ATA) and methods to complete ATA for use in a ca-MST. The section will conclude with a brief discussion of methods to control over-exposure of items in different testing mediums. The three topics discussed in this section are important because these topics discuss methods of determining not only the location of information for the item pool itself but also once the item pool is developed methods of Automated Test Assembly which helps the individual modules and panels as a whole unit become interchangeable with one another. Item exposure and precision are completely dependent on the item bank and affect such issues as test cheating and precision of scores.

Methods of Determining the Central Location for the Item Pool

One of the few resources on item pool maintenance comes from Breithaupt, Ariel, and Hare (2010) who describe assembling an inventory for use in a ca-MST. Different types of testing programs will require unique types of item pool maintenance. For example, ca-MST and CAT have different needs in terms of the location of information within the item pool. The next section will compare the item pool maintenance strategy for ca-MST to the item pool maintenance for other tests such as CAT (Veldkamp & van der Linden, 2010).

Research on this topic has suggested that item pool maintenance based only on adding new items to a pool is not sufficient in order to uphold standards of item pool

quality over time (Ariel, van der Linden, & Veldkamp, 2006). Breithaupt (Breithaupt et al., 2010) argues that there is a natural tension between the goal of creating a large enough item pool, with relevant item exposure controls that are equally distributed across items and cost reduction and efficiency in item pool creation. In this research, Breithaupt develops and describes the "Steady-State Model). In this model, an ideal bank is described. The authors argue that the cost of producing complex performance items adds to a major problem of maintaining a large high quality bank. As such, balancing cost with a minimally tolerable item exposure is important to a testing program that requires these types of items. The first consideration in determining what the authors claim, as an ideal minimum-size bank is content. In order to determine how much of a particular content, the length of a test form, and the proportion of test content or skill area is known, the number of items covering that particular content can be derived.

Another consideration that needs to be taken is the retirement or disclosure rules that govern the assessment. In order to determine how often to retire items, the topic needs to be clearly monitored. For example, Breithaupt (Breithaupt et al., 2010) mentions that the lifespan for case studies in a clinical (medical) context should have a shorter lifespan because of the fast paced evolution in relevant content, while items used for an algebra assessment/classroom assessment may be able to be retired at a far less rapid pace.

One benefit of a ca-MST is that a predictive model can be directly mathematically calculated to determine the proportion of candidates will see any given item. This probability depends on the number of panels, as well as, the number and position of the

models within a panel. All that is needed is the probability of a student seeing a particular panel and the proportion of students that are routed to the module.

Veldkamp and van der Linden (2010) take a slightly different approach to the maintenance of item pools. This research focuses on the use of traditional item pool blueprints and content blueprints to help identify shortages of content/statistical concerns within an item pool. This paper uses a Shadow-Test approach. Shadow tests are full-length tests that are optimal at the current ability estimate that meet all test specifications and contain all items given to the test-taker. Then constraints are placed on each form that minimize costs and constrain things at the test level such as test information, length, number of stimuli, and content constraints. One must put careful consideration into defining what Veldkamp (Veldkamp & van der Linden, 2010) calls "the design space" (p. 234). This differs from Breithaupt's methods of maintaining an item bank, however, is a useful method for maintaining proper item demands in the item bank.

Ensuring proper levels of information (precision) and security (item exposure) should always be balanced with the cost of producing enough items in the item pool. Dallas et al. (2012) illustrated that not only does exposure and security have a balance with cost but also with each other. Decisions made about which ca-MST configuration chosen can ultimately decide the proper amount of exposure to obtain minimum amounts of precision and cost. Ultimately, the testing purpose should always guide decisions about where precision and exposure are key and where precision may not be so important. For example, a certification test that has a single cut at $\theta = 0.00$ may not need a great deal of items that can give great amounts of precision at $\theta = 2.00$.

Aside from exposure another important consideration for any assessment system is the location of the information and the scale precision. The location and amount of the information across the scale can mean the difference between a successful and unsuccessful testing program. Too little information and precision is lost, too much information and money is spent on something that might never be used. Also, if there is plenty of information but the information isn't in the location where it is needed, the results can leave developers with a major problem. The information function has applications in not just test construction, but item selection, assessment of precision of measurement, comparison of tests, determination of score weights, and comparison of scoring methods (Hambleton & Swaminathan, 1985).

The test information function can generically be written as

$$I(\theta, \hat{\theta}) = \sum_{i=1}^{n} \frac{P'(\theta)_i^2}{P_i(\theta)Q_i(\theta)} \quad \text{(Eq. 4)}$$

The amount of information is influenced by the quality and number of test items in the pool/on the form/in the ca-MST (Hambleton & Swaminathan, 1985). The items that will provide the most information will have steep slopes (high a-parameters) and low item variances. Item information is summative and independent of other items. There are particular strategies such as Assessment Engineering (Luecht, 2006) and Evidence Centered Design (Huff, Steinberg, & Matts, 2010; Mislevy, Almond, & Lukas, 2003) that can aid item writers in developing items with these features. The optimal method of doing this is by taking exemplar items or items with high a parameters and modeling them in an item shell type format. Methods such as Assessment Engineering's Task

Modeling (Luecht, Dallas, & Steed, 2010) can help bring item shells into a common language, as well as, allow for multiple instantiations of task models called templates to create several items with similar statistical qualities.

Several empirical studies (Bejar, 1991; Bejar et al., 2003; Embretson, 1999; Luecht, 2011) have tested whether item shells or items from task models would be statistically feasible. Bejar et al. (2003) examined the use of on-the-fly item generation using item shells. The study used item models from the quantitative section of the GRE in an adaptive testing (CAT) setting. The study sought to test whether or not items could be generated that contained a great deal of isomorphicity or equivalent content and psychometric to the current GRE. The simulation study showed mixed results. The study used four scenarios. First, no isomorphicity, this scenario contained each isomorph produced one and only one item and item parameters were known. The rest of the simulations had several isomorphs generated under an item shell with different levels of variance in the b parameters. The worst-case scenario had the most amount of variance in the b parameters, while the best-case scenarios contained little variance in the b parameters under an item shell. Under high levels of isomorphicity the standard error of measurement, or the inverse of information, was the same as the no isomorph condition; however, as variance was increased among the isomorphs precision was lost. The field study; however, showed a correlation between the standard GRE and the on-the-fly GRE created from item shells to be .87. This correlation was in line with the general test- re-test correlation found between normal administrations of the GRE. This indicates that the item shells function just about as well as typical items. With this in mind, the use of item

shells on the best items may prove to be a reasonable method for using current item pools to increase the quality of the item pool to obtain the most amount of information. This research is important to this study because item shells can be combined with Automatic Item Generation (Gierl & Lai, 2011) to produce thousands of items for use in the item bank. While not yet operationally used, item shells and AIG are cutting edge technology that could represent the future of item pools.

Methods for performing Automated Test Assembly (ATA). The 3PL model is specifically being used to develop items for the item pools for this paper. The item information function for the 3PL model is as follows

$$I(\theta, \hat{\theta}) = \frac{D^2 a_i^2 q_i}{p_i} [\frac{(p_i - c_i)^2}{(1-c_i)^2}] \quad \text{(Eq. 5)}$$

Since item information is summative, the test information function would just be to sum across the items and theta scales. In order to build the TIF's, this paper uses Automatic Test Assembly using an ATA engine. Most ATA engines seek to minimize or maximize functions to fit certain constraints on either content or information such as this target which minimize the gap in Information across the θ scale:

$$\mathrm{T}(\theta) - \sum_{\mathrm{i}=1}^{\mathrm{I}} \mathrm{x_i}\, \mathrm{I_i}(\theta) \quad \text{(Eq. 6)}$$

$\mathrm{T}(\theta)$is the target and

$$\sum_{\mathrm{i}=1}^{\mathrm{I}} \mathrm{x_i}\, \mathrm{I_i}(\theta) \quad \text{(Eq. 7)}$$

is the summative item information across the θ scale or the total test information currently in the exam.

Luecht and Nungester (1998) argued that there are two ways to perform the ATA for a ca-MST. The first called "top-down" assembly requires the specification of content constraints for the total test along all possible pathways thru the panel. The ATA engine would then optimize a model that would aggregate item content across the stages to satisfy total test level constraints. Also, TIFs would be generated for the "most likely" pathways for an examinee.

Luecht and Nungester (1998) also described another approach called the "bottom up" approach looks at specific routes that a low proficiency examinee, moderate proficiency examinee and a high proficiency examinee might take through the panel. In separating these out modules become mixable between panels as a module in the moderate position can be switched between stages or between panels as modules now become separate. Separate target TIFs also would be created for each module.
One important decision that a developer makes in the ATA process is the level of exposure that is tolerable for the assessment. Luecht (2000) expanded the design framework and automated test assembly mechanisms needed to implement ca-MST in practice. Building on the original CAST model, he demonstrated how to build forms using ATA, while simultaneously focusing on critical item exposure issues. Luecht argued that the advantage that ca-MST has over CAT is due to using robust average test information targets for the modules, ca-MST designs evenly distribute the "best" or highest a-parameter items across the panels. One drawback CAT has is that since the

algorithm is going to pull the item with the most information, high a parameter items will always have the most information. In addition, because we do not re-use items on panels, item exposure can be seen as consistent across a stage. For example, if there is a 1-3-3 design with 10 panels at each module, the probability of any item being seen is the probability of being routed to low medium or high, 0.33, times the number of random panels at each stage 10, or 0.033. Lastly, exposure can be empirically calculated as mentioned in the last step. Assuming we control the item to be on 1 and only 1 panel or limit an item to a particular number of panels, we can easily calculate the maximum exposure rate for any given item. These decisions can even be incorporated into the Automated Test Assembly process as a constraint on exposure. Also since we know which panels will be active for any examinee, corrections could potentially be made in advance to remove items with excessive exposure, content has become outdated, or the content has become exposed due to potential test cheating.

Luecht and Burgin (2003) described a solution to building TIF targets called conditional information targeting (CIT). Luecht argues that this strategy will be generalizable to having multiple routes and stages. The CIT strategy relies on three values called "posts." The left, center, and right posts each have different roles in generating the TIF functions. In general, the posts are related as follows: $\theta_L < \theta_C < \theta_R$. The center post is always fixed and controls the proportion of the population sent to the left and to the right. The left and right posts are used to move the targets for the maximum amount of information the modules. The TIF strategies explained below are just generalizations of the CIT principles.

Luecht and Burgin (2003) found that the targets could be hit, however recommended fixing one of the outer posts helps constrain an otherwise challenging solution, where aligning the intersection to the center post is the primary criterion. This is particularly challenging when the supply vs. demand gap is particularly narrow. In general a lack of convergence to a ATA solution can cause the intersection to move or not be met such as in Luecht and Burgin (2003) when one of the solutions ended up with a θ_C = -0.46 when the criterion was set at -0.524. Luecht and Burgin admit that there may be an issue with more than two modules as the research only included a 1-2 design. There may also be some concern with further than two stages, as some people may take auxiliary routes and the CIT strategy only works for the primary routes. While the thought is that particular auxiliary routes should balance out or the people going from a moderate module going to easy should be the same as the people going from an easy module to a moderate module, this should be empirically researched.

Several exposure control methods exist for CAT and LFT. As previously discussed the majority of exposure control methods rely on the Sympson-Hetter procedure (Sympson & Hetter, 1985). There are other types of exposure controls that have also had success such as Chang and Ying's (1996) a-stratified exposure controls. Way (1998) classified all exposure control procedures into two main categories: randomization selection procedures and conditional selection procedures. Randomization procedures select several items near the optimal level of maximum information from which one item is then randomly selected for administration in a CAT. Conditional selection procedures have preset exposure parameters that meet a pre-selected maximum

exposure rate. Dodd and Fitzpatrick (2002) note that if the ability distribution is changing the conditional selection obtaining the exposure parameter is a nontrivial process. This section will discuss a couple of the more popular exposure control methods.

The most commonly used conditional selection procedure is the Sympson-Hetter (Sympson & Hetter, 1985) technique. Ultimately, Sympson-Hetter limits the exposure of items that have been exposed greatly before. This ensures essentially a maximum exposure rate because the higher the frequency of administration in a CAT the smaller the exposure-control parameter that the item will have.

Another commonly used conditional exposure control method is the a-stratified (Chang & Ying, 1996) method. In this method, when the CAT algorithm is most uncertain of the examinees' ability, lower discriminating items are administered. As the CAT progresses, the examinees' ability becomes more known and higher discriminating items are administered. This method falls into a joint approach to exposure control. The items within each strata (low and high discrimination) are randomly distributed to an examinee. As mentioned earlier there are some questions as to how many strata should be created and what the strata should look like. Overall, these types of can help hit exposure goals in a CAT; however, unlike CAT, ca-MST will have known item-exposure characteristics and these item exposure characteristics are entirely dependent on the item pool and decisions made about the ca-MST configuration (see: Dallas et al., 2012).

These topics are important in developing what ca-MST configuration that will work best for your assessment. The item pool and the information and shape of the item pool is extremely important in determining which ca-MST configuration will serve the

test purpose best. This section described a couple of the available item pooling techniques for ca-MST. Followed by a discussion of a couple of the different type of ATA procedures found in the literature. Finally, a couple of the popular exposure techniques used in CAT and could potentially be used in a ca-MST. These topics inform the final research question of supply versus demand. Ultimately, the ca-MST will only be as good as its supplies (information in the item pool) versus the demands placed on it by the choices made about configuration and ca-MST options (such as the # of items in a module or item re-use rules.)

Routing and Scoring Methods for ca-MSTs

There are varieties of ways to route examinees from stage to stage. CAT procedures usually provide the item that optimizes item information at the examinees $\hat{\theta}$ estimate; however, ca-MST's can be a little bit more flexible in the method that routing is performed. As noted in Chapter I, four types of routing and scoring are considered in this study: (a) routing using Bayesian a maximum module information criterion using Bayesian mean (i.e., *expected a posteriori* or EAP) scores; (b) routing to achieve a balanced exposure of modules within any ca-MST stage using EAP scores and cut scores along the θ distribution; (c) routing using number-correct (NC) scores, but with the underlying module information functions designed routing the examinees using a maximum module information criterion; and (d) NC-scoring with routing set to achieve uniform balance of module exposure within a stage.

Very little has been done directly addressing the effects of the consequences of a particular routing strategy. Yen (1984) investigated the use of a NC scoring method on a

LFT. Yen's research used a look up table where the NC score was used to find an examinee's estimated trait level using the 3PL model. This was done by using an approximation to the compound binomial distribution (Lord & Novick, 1968). Yen found that NC scoring produced similar results except in the lower two quintiles of the scale. In addition, Yen found that there was a 10% increase in the average error of measurement for the $\hat{\theta}$ distribution. Stocking (1996) proposed a method for scoring fixed length CATs based on a NC raw score. While a LFT contains multiple raw score options with approximately the same mean difficulty for a form, CAT contains about the same raw scores with the only difference being the mean difficulty of the items taken by the examinee (Dodd & Fitzpatrick, 2002).

In terms of NC scoring on ca-MSTs, Schnipke and Reese (1999) used NC scoring in their simulation study in which they used an approach that attempted to figure out at which NC value was the mean squared error lowest between pathways (low to medium, medium to high). Dodd and Fitzpatrick (2002) also offered up a routing method that is similar to the information based routing method used in this study. Dodd and Fitzpatrick's method involved computing NC ability estimates and then selecting modules based on information at that estimate. Schipke and Reese (1999) found that selecting testlets on the basis of information worked better than selecting testlets of a given average difficulty based only on the raw score of the previous testlet.

Luecht, Brumfield, et al. (2006) tangentially dealt with routing in a study examined some key practical issues/development considerations. They specifically examined how a 1-3-3 ca-MST (then called a Computer Adaptive Sequential Test or

CAST) would fit the Uniform CPA Examination. They chose to do number-correct routing to route examinees to the next stage. While the authors did not compare number correct routing to other routing methods using IRT, they did bring up an important technical consideration if using number-correct routing. There are two potential methods for locating the routing points on the ability scale.

The first method described by Luecht, Brumfield, et al. (2006) included the use of IRT estimates to develop cut scores for the panel. One would obtain number correct scores for specific cuts and use those number correct decisions to route examinees through the panel. These number correct scores could either be cumulative or based solely on the previous module. The IRT cut score formula would be calculated for a set decision point by summing the probabilities of a correct answer across items in any of the IRT models given the set of item parameters:

$$X_{cut} = \sum_{j=1}^{k} \sum_{i=1}^{n_j} P(\theta_{cut}; \xi_i) \qquad \text{(Eq. 8)}$$

The other method of routing forces a defined population interval to particular routes. This method can be used more as a policy decision to force relative proportions of examinees to each of the primary routes. This would be accomplished by finding the ability points that would correspond with the 33rd and 67th percentiles of the cumulative distribution. Assuming a normal distribution or ($\mu = 0$, $\sigma^2 = 1$) the routing points would be $\theta_{LM} = -0.44$ and $\theta_{MH} = 0.44$.

These two methods are somewhat similar to the methods used in this study. One type of NC routing focuses on efficiency and precision routing the examinee to the route

that is maximally informative given prior performance on the assessment. The other type of NC routing focuses on exposure routing examinees to routes based on a designed interval to protect the modules from being over exposed.

No matter which method is chosen number correct scoring rules can be set up. They can be individualized for each panel (i.e., if equating differences exist). Once these rules are set up, the only IRT needed is the end score ability estimate. The payoff with using this method is that no longer will testing centers have to perform IRT calibrations on the fly in order to route examinees. To date, little research has been done confirming that number correct methods work equally well to IRT methods of scoring.

The other methods to score examinees would be to use $\hat{\theta}$ to score examinees. $\hat{\theta}$ is computed by using the underlying Item Response Theory model, most likely the three parameter logistic item response model (3PL). The most likely $\hat{\theta}$ could be found by computing the likelihood across the ability scale by computing the likelihood function on θ. Examinees would be given more credit for correctly answering items that are either highly discriminating or highly difficult. IRT has two potential methods of routing: routing by the module that will provide maximum information at the examinees current $\hat{\theta}$ proficiency level or have specific $\hat{\theta}$ cuts set up for advancement to the next module.

The first IRT method that is under consideration is used primarily in CAT research. Most CAT routines select an item that is maximally informative to the users $\hat{\theta}$. This method would route students ultimately the best as research has shown (Hambleton & Swaminathan 1985; Lord, 1980; Luecht, 2000) that when item difficulty matches ability both efficiency and validity are maximized. In this scenario, $\hat{\theta}$ would have to be

calculated at least at the end of each module and at every routing decision point, additionally, the routing mechanism would have to be the cumulative method of scoring, which may reduce the total amount of adaptivity in the design; however, ultimately this method has traditionally proven to produce scores that are the most efficient and precise. One drawback of this method is that proportions of the population are not as easy to control. Information targets can be set up to roughly route examinee proportions to roughly desired conditions; but $\hat{\theta}$ does not have the strict controls that number correct routing or θ cuts have over controlling the proportions of people whom see each module.

The hybrid between number correct routing and max information routing is the last of the routing strategies and that's providing $\hat{\theta}$ cuts and routing people according to those cuts. This method is similar to the information targeting method in that IRT is performed at each module and the $\hat{\Theta}$ estimate is directly tied to the module that the examinee is sent to. In addition, it also offers the strict population control procedures that can be found in number correct routing.

One of the primary differences between number correct scoring and $\hat{\Theta}$cut scoring is simply the method of calculating the actual cuts. Recall that in number correct scoring IRT true scores were calculated by simply taking the item parameters, ξ and the cut score for θ_{cut} and summing the probabilities of a correct response at that cut score. The difference is no longer do those true scores need to be created; the θ cuts will suffice in determining the cut scores to route. The scoring can either be done cumulatively or separately for each stage; however, in order to obtain the most precise estimates for $\hat{\theta}$, the cumulative approach to scoring should probably be used.

A bit of work has been done in using the Test Characteristic Curve (TCC) or a curve of the expected raw score for any given θ. This research in some way related to using the raw score to approximate back to the θ scale. Thissen and Wainer (2001) mentioned the idea of using the TCC to score tests (i.e., using "sum scores") for the North Carolina EOG Mathematics Test (Williams, Pommerich, & Thissen, 1998).

Kolen and Tong (2010) evaluated the effects of choice of estimator on score distributions while considering the effects of the prior and test length. The paper compared TCC scoring, MLE scoring, EAP scoring and a variant of EAP scoring for a summed score estimator. The findings from this study have two major implications for number correct scoring. First, Kolen and Tong found that the score distributions were more variable for the MLE/TCC scoring functions as compared to their Bayesian counterparts. Second, in practical situations, number-correct scoring is more likely easier to understand for users of the test than IRT scoring (Kolen & Tong, 2010).

While this study isn't explicitly using TCC scoring, the routing of examinees is completely done using a number correct method. Final score estimates will be both MLE and EAP IRT estimates and not a TCC score. That said, using number correct (summed score) estimates to route is not a brand new idea and has been done in previous research.

While more has been done in this area, there is a routing method that this paper did not consider. This method was Wald's (1947) sequential probability ratio test which was explored further by Luecht et al. (1996). This method was not seriously considered, however, as suggested by a comment made in Luecht and Nungester (1998) to the effect that the SPRT was considered valuable only in a mastery testing context where

improving the quality of ability estimates was not a major concern. In the Luecht et al. study the ability estimates are of concern and as a result the SPRT is not an appropriate routing method. Another method of routing /scoring that was not tested is the method developed by Luecht and Burgin (2003) which is a hybrid method of scoring which routes the examinee to the module most informative, however, also keeps exposure in mind in the process.

Summary

The development of ca-MST has begun to take hold on the testing industry as more and more companies wish to integrate cutting edge technology with cutting edge measurement methods. This notion should sound the call for renewed research into ca-MST and focus the bulk of the research on specific technical considerations that should not be overlooked.

Developing cutting edge item pool development strategies that maximize precision where it is needed most and routing examinees in the most efficient way in terms of efficiency and exposure are keys for any testing company. Companies can save money if item pool demands can be known in advance and dealt with in a reasonable manner. More research needs to be done to determine the best types of information structures for ca-MSTs and how item pool maintenance can be done in the most efficient manner possible.

This literature review shows the gap in the assessment the effect of routing strategies as well as the history of measurement in the field of ca-MST. These

considerations will go a long way in determining the future of not only ca-MST assessments but in assessment as a whole.

This literature review highlights the importance of the research questions addressed in this paper. Routing and scoring methods for ca-MSTs are not well established in the literature. This dissertation represents the first empirical study of routing and scoring conditions testing them purposefully against other routing and scoring conditions. While there is nothing necessarily new about the routing and scoring conditions themselves, the use of these conditions within a ca-MST is untested. Lastly, a discussion of item pooling and item exposure methods highlights the final research question. The final research question investigates the notion of striking a proper balance between available psychometric information and the demand that the configuration and other ca-MST options places on the item pool. Creating an item pool should not just be a simple repurposing of available items from another item bank and making a ca-MST out of those items. Item pools for ca-MST should be constructed with some forethought and constructed to meet the needs of the desired ca-MST configuration as opposed to letting the item pool control the configuration options or performing suboptimal measurement (Dallas et al., 2012).

CHAPTER III

DATA AND METHODOLOGY

This chapter provides an overview of the data generation and simulation process and the study design conditions. A brief summary of analysis procedures is also included. Those procedures are more fully covered in Chapter IV.

Data Generation and Simulation Process

The research questions outlined in Chapter I are addressed using a large-scale simulation study covering 36 conditions, with multiple data set replications per condition, and simulated examinees serving as their own controls across four within-person ca-MST routing and scoring strategies. Details about the 36 design conditions and four routing/scoring strategies are covered in the next subsection.

Each simulation is carried out in four parts. Part 1 is the data generation and calibration phase. An item pool is created specifically for each study design condition. These item pools were simulated to be roughly 1.5x the information needed for a paper-pencil fixed length assessment. The "known" item parameters are used to generate a calibration response data set using procedures outline below. The examinees in the calibration sample are assigned a random set of items—essentially a LOFT form—matching the corresponding panel test length (e.g., for the 1-3 panel configuration with 8 items per stage, examinees in the calibration sample would be assigned a 16-item LOFT form. These calibration samples will be sufficiently large to ensure that at least 500

examinees see any particular item and that sufficient overlap or connectivity in the data set exists for calibrating the entire item pool to a common IRT θ scale. The calibration data set will be calibrated using BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 2003). All item parameter estimates for the item pool must converge. This calibration method simulates a reasonable approach to calibrating an item pool based on field trials or embedded field-testing of the items.

Part 2 is the panel assembly phase. Here, ten replications of the panel configuration are assembled from the item pool, without item overlap across panels, using the calibrated item statistics from Part 1. Appropriate routing and scoring tables are also created as part of the assembly process. All of the assembly is carried out using a well-established automated test assembly (ATA) heuristic for ca-MST designs (Luecht, 2000). A variant of the CASTISEL version 1.5 software, called ATA Panel Builder (Luecht, 2013) is used.

The targets for the ATA are developed using R-programming (2013). First, the location and amount of maximum information is computed. For this task, the TRUE (generated) item parameters for this task. Next, based on the maximum information per item, the items will be sampled into *S* bins where S = the number of stages (2, 3, 4, or 5). The size of each bin should be proportionally equal to the sum of the module sizes at each stage. For example in a 1-2-3 design with 10 item modules, the item pool would have 720 items (1.5x the demand). Bin #1 would correspond to stage 1, which has only one module. As a result, we would sample 1/6th of the item pool and assign them to Bin #1, stratifying by max. amount of information to ensure a somewhat uniform spread of

information across the modules, regardless of the mean of the difficulty in the item pool itself. Bin #2 or the second stage items would have 1/3rd (2/6ths) of the items in the pool and the remaining ½ (3/6ths) of the items would end up in Bin #3 for use in the third stage. Next, the items in each bin will be sorted by location of maximum information. These items will then be split into "piles" corresponding to the number of difficulty levels (modules) at each level. For example using the previous example using a 1-2-3 ca-MST configuration, Bin #1 would need no sort or split needed because it only has one level of difficulty. Bin #2 would sort and split the items into a "moderately easy" and a "moderately hard" pile. Bin #3 would need to be sorted and split into three piles: an easy, medium and hard pile.

Next, the average item information across all items within each pile in each pin will be computed at 41 quadrature points. The average item information across each bin then becomes the TIF targets for the ATA procedure. The ATA procedure uses the Normally Weighted Absolute Deviation Heuristic (NWADH; Luecht, 2013; Luecht, 1998), or a variation of a greedy algorithm. This algorithm is useful because it always converges to a solution, even if there is not adequate information in the pool. In addition, when there is a lack of information in the pool the algorithm will do the best job it can in building parallel forms. The NWADH procedure updates the overall information after selecting an item and adjusts the "best" item (or the item that meets the biggest gap) after each selection.

Part 3 generates a simulation data set using the original item parameters from Part 1 (not the calibrated statistics). The data generation is carried out for the entire item bank

to ensure that the same response generating mechanism is applied for a particular simulated examinee, regardless of which panel or routing/scoring mechanism is applied in Part 4, below. Twenty thousand examinees are generated per simulation data set to ensure that simple random assignment of panels to examinees will result in approximately 2000 examinees being administered any particular panel.

Part 4 is the ca-MST simulation phase. This phase uses the generated simulation data sets from Part 3 and the pre-assembled panels from Part 2. These simulations are carried out by caMST Simulator (Luecht, 2013). In this phase, the ten pre-assembled panel replications are randomly assigned to each simulated examinee in a simulation data set. The same panel is then essentially administered four times, once per routing and scoring strategy (described in Chapter I and below). The estimated item statistics from Part 2 are used for all routing and scoring, even though the response data are generated using the original item pool statistics from Part 1. This process allows item parameter estimation errors to potentially impact the accuracy of scores, in addition errors attributable to the various routing and scoring mechanisms used. Since the four routing and scoring mechanisms are applied to the same response string that each examinee, the examinees essential serve as their own controls for comparative purposes.

The calibration data sets (Part 1) and ca-MST simulation data sets (Part 4) are rather straightforward extensions of data generation methods used in most IRT-related research. Here, we consider an item pool consisting of I items with parameters, $\xi_i=(a_i, b_i, c_i)$, $i=1,\ldots,I$. The item parameters are sampled from target distributions of: (a) item discriminations, $\ln(a)\sim N(\mu_a, \sigma_a)$; ($b$) item difficulties, $b\sim N(\mu_b, \sigma_b)$; and (c) lower

asymptote parameters, $c \sim B(\mu, \sigma)$, where *N* denotes the Gaussian normal distribution and B denotes a beta distribution. The moments of the sampling distributions are specifically manipulated to achieve the study design characteristics outlined above.

These sampled item pool parameters are used in the IRT three-parameter logistic model (3PLM) to generate dichotomous (1 = correct, 0 = incorrect) response data for the entire pool of items, given a sampled proficiency score, $\theta \sim N(\mu_\theta, \sigma_\theta)$. The 3PLM can be written as

$$\Pr\left(u_i = 1 \middle| \theta, \xi_i\right) \equiv P_i = c_i + \frac{1 - c_i}{1 + \exp\left[-Da_i\left(\theta - b_i\right)\right]} \qquad \text{(Eq. 9)}$$

For each person-by-item interaction, we can apply Equation 3.1 to obtain an item "true score," P_{ij}, using ξ_i and θ_j (i=1,…,I items and j=1,…,N examinees). We can next generate a uniform random real-valued number within the interval $0 \leq \pi_{ij} \leq 1$. If $P_{ij} \geq \pi_{ij}$, we set the scored item response to u_{ij}=1, otherwise u_{ij}=0. The process is repeated for each simulated examinee on every item in the pool. However, only those responses actually administered in Parts 1 and 4 are used in scoring. The full-pool item generation is primarily used for data management convenience, requiring a total of only 540 complete responses data sets for Parts 1 and 4.

Study Design Conditions

Thirty-six conditions are considered in this study. The design factors cover two general categories: (a) ca-MST design configurations, and (b) item pool characteristics. There are 6 plausible ca-MST design configurations consisting of two module-lengths

and three panel configurations. The module lengths are 10 and 20 items. The ca-MST panel configurations include: (a) a 1-3 two-stage design; (b) a 1-2-3 three-stage design; and (c) a 1-2-3-4 four-stage design. These choices appear to reflect an adequate range of adaptation for two- to four-stage panels and expand on the work done in previous studies (e.g., Jodoin et al., 2006; Wang et al., 2012).

Six characteristics potential item pools are fully crossed with the ca-MST. The choice to focus on item pool characteristics, rather the spread of measurement information within particular modules or panels emphasizes a realistic condition of test assembly. Item pools are not infinite resources. By focusing on the item pool characteristics, we force the panel assembly process to deal with the realities of a specific item pool and then generate a number of modules that contain similar targets. The generated pools are further limited to being 1.5 times the number of items needed to create twelve parallel panels under one of the six ca-MST design configurations described above. Using twelve parallel panels is a reasonable number of panels that a test developer might construct to ensure that random assignment of panels would deter any cheating/gaming of the ca-MST system. In addition, these panels serve a method of replication in case a particular panel has something go wrong in the simulation. This seems to be a reasonable restriction that might mirror practical item inventory planning. The six sets of item pool characteristics are determined by manipulating the means of the item discrimination parameters and the means of the item difficulty parameters. Details on the IRT model used and how those item discrimination and difficulty parameters are expressed in the model are explained in the next subsection.

If we consider a distribution of examinee proficiencies drawn from a unit-normal distribution, $\theta \sim N(0,1)$, we can move the statistical measurement information of an item pool relative to that examinee distribution and the purpose of the test. Three levels of average item difficulty (μ_b) are used to simulate the item pools: (i) item pools comprised primarily of easy test items ($\mu_b = -1\sigma_\theta$); (ii) item pools composed of primarily moderate difficulty items—that is, an item pool well-matched, on average, to the examinee population mean ($\mu_b = 0$); and (iii) item pools comprised primarily of difficult items ($\mu_b = +1\sigma_\theta$). All of the item pools have an average b parameter variation of 1.0. Using 1.0 represents, moderate variation in the item pools ($\sigma_b = 1.0$), and simulates item pools well matched to the variance of examinees in the population. Most of the other research studies use 1.0 as the standard deviation for the difficulty parameter (Xing & Hambleton, 2004; Jodoin et al., 2006).

Two levels of average item discrimination were also consider: (i) moderately low discrimination, μ_a=.6, representing a plausible amount of discrimination for many types of certification and licensure tests administered to highly able, homogeneous samples of examinees; and moderately high discrimination, μ_a=1, a value compatible with a number of large-scale achievement and college entrance examinations. Most studies included an average discrimination parameter of 1.0 (e.g., Xing & Hambleton, 2004). Note: unrealistically high discrimination values are sometimes included in simulation studies. However, those levels rarely contribute anything useful to an operational assessment system because levels of average discrimination above 1.2 or so are almost impossible to consistently maintain given even very good item writing guidelines and training of item

writers. In fact, a recent study by Paap and Veldkamp (2012) examined the effect of a testlet model and the relationship between the discrimination and difficulty parameters. The average a's in this study ranged from 0.4 to 0.9.

The complete set of crossed design conditions is summarized in Table 3.1. Taking the product of the "Totals" row values gives the 36 total conditions noted above. Each of these conditions will be replicated 10 times for a total of 360 item pools in the study.

Table 3.1

Summary of Study Design Conditions

	Panel Designs		Item Pool Characteristics	
	Module Length	ca-MST Configuration	Average Item Difficulty	Average Item Discrimination
Totals	**2**	**3**	**3**	**2**
Level 1	10	2 Stage: 1-3	Easy, $\mu_b = -1$	Low, $\mu_b = .6$
Level 2	20	3 Stage: 1-2-3	Mod., $\mu_b = 0$	High, $\mu_a = 1$
Level 3	–	4 Stage: 1-2-3-4	Difficult, $\mu_b = 1$	–

Note: Each of these 36 item pools will be replicated 10 times.

As noted in Chapter I, four types of routing and scoring are considered in this study: (a) routing using Bayesian a maximum module information criterion using Bayesian mean (i.e., *expected a posteriori* or EAP) scores; (b) routing to achieve a balanced exposure of modules within any ca-MST stage using EAP scores and cut scores along the θ distribution; (c) routing using number-correct (NC) scores, but with the underlying module information functions designed routing the examinees using a

maximum module information criterion; and (d) NC-scoring with routing set to achieve uniform balance of module exposure within a stage. As previously mentioned, these four routing and scoring methods directly address the gap in the literature on number correct scoring versus traditional scoring practices.

For maximum information routing with Bayesian scoring, we compute the module information functions (MIF) at the current EAP estimate and choose the module with maximum information (Willse et al., 2012). For NC routing with the maximum information criterion in place, we determine the intersection of any pair of adjacent MIFs and compute the rounded expected NC score using all items completed through prior stage along designated ca-MST routes; that is, $\text{Round}[E(NC)=\sum_i(\theta_{intersection})]$. The rounded expected NC value serves as a boundary score for routing purposes. For uniform, balanced module exposure, we divide by the number of modules at any one stage and determine the corresponding θ (or estimate) corresponding to the area within each of the regions of the cumulative normal distribution. For example, for three modules at a particular stage, we divide by three and determine the cut points at the 33rd and 67th percentiles of the cumulative normal distribution to route approximately equal proportions of the sample to one of the three modules. When IRT Bayesian scoring is used, the routing is based on EAP estimates of θ. When NC routing is used, we determine the expected NC cut points using the $\text{Round}[E(NC)=\sum_i(\theta_{cut})]$ transformation. This description also helps to highlight that the primary difference between the IRT-based routing and scoring and the NC approach is that the latter depends on the rounding

function. Its advantages lie in simplifying the operational routing and scoring mechanisms and IRT-related data needed to support ca-MST in real-time.

Data Analysis

Once $\hat{\theta}$ was calculated for every examinee, residual statistics will be calculated and reported for each examinee within each design condition. For each $\hat{\theta}$, the residual for any jth examinee can be calculated as $e_j = \hat{\theta}_j - \theta_j$. The residuals serve as a measurement of error and the expected value of the residuals is expected to be zero. The equation for the mean of the bias is

$$\mu_e = E_j = \frac{\sum_{j=1}^{N} e_j}{N}. \quad \text{(Eq. 10)}$$

When the mean of the residuals is non-zero, however, the estimates of $\hat{\theta}_j$ is said to contain bias. Pine (1977) and Hambleton and Swaminathan (1985) described item/test bias as the same underlying ability (measured by θ) having the same probability of getting an item correct. Most commonly bias serves as a measure of finding Differential Item Functioning (DIF); however, since there was only one group in this study, ultimately bias takes on the meaning of do all examinees with similar abilities perform at the same level. In other words, does the choice of a ca-MST, model selection (e.g., two parameter IRT model, three parameter IRT model, or routing method (number correct, IRT $\hat{\theta}_j$, or max information IRT routing) cause groups of individuals with the same $\hat{\theta}_j$ to be misestimated in a similar manner? For example, if somehow the number correct scoring method caused everyone in a range of ability from -1 to 1 to be misestimated in a similar direction, the estimates could contain bias. If it is the case that the number correct

method leads to greater amounts of bias than the IRT methods of ability estimation, the number correct method probably should not be used operationally.

The other outcome observed was also related to the residual error. Root Mean Square Error (RMSE) was defined in this paper as

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{j=1}^{N} e_j^2}{N}}. \qquad \text{(Eq. 11)}$$

RMSE is the square root of the Mean Square Error (MSE). The expectation of RMSE in a non-biased distribution is expected to be 0. Luecht (2011) notes that MSE is the sum of the squared errors, RMSE subsume bias and error variance into a single measure. It should be noted that the error variance is the variance of the residuals around the mean of the residuals. The MSE should be similar to the empirical error variance and should only be different by the amount of bias or the mean of the residuals. In this study, the RMSE will be used to compare the true theta to the estimated theta and how well the ATA resulting MIFs matched the target MIFs.

Review of Research Questions

These analyses are directed at answering the three research questions in this study. Ultimately, the goal will be to show how differing both the ca-MST conditions and the within-person routing mechanisms affect precision of estimation of ability.

1. *How do the four routing and scoring mechanisms impact the recovery of simulated proficiency scores under a variety of ca-MST conditions?*

The two statistics as earlier described will be computed and then graphics will be developed to illustrate the overall effect of the routing conditions under a variety of ca-

MST conditions. Ultimately, this is similar to examining just the main effect of the routing conditions, as routing condition accuracy is a major concern to test developers in establishing validity. In addition, conditional residuals will be calculated and plots or tables of route residuals (bias and RMSE) will be calculated. The results from this question will directly address the gap in the literature about the possible use of NC scoring in an operational framework.

2. *Under what conditions do various ca-MST design configurations work best from a psychometric scoring-accuracy perspective with respect to the four routing and scoring mechanisms as well as the various item pool characteristics?*

Graphics will be developed to illustrate which design conditions do various ca-MST conditions work best from a scoring accuracy. This question examines the interaction between configuration and routing condition controlling for the other study conditions. The graphics from this question will examine the potential issues that can arise as a result of certain item pool characteristics and the potential interactions between the ca-MST configuration and psychometric scoring accuracy.

3. *How do various item pool characteristics impact the quality of ca-MST test forms from a psychometric perspective?*

The two previously discussed statistics will be calculated and then graphics will be developed to investigate this question. This research question will be answered via graphics that directly address how item pool characteristics affect the quality of ca-MST forms from a psychometric perspective. This question directly addresses the supply vs.

demand question from Chapter I. The basic relationship between supply and demand might impact the ca-MST configuration decisions that a testing company might make. For instance, if the item pool won't support the decisions made it may hurt the testing program in terms of the psychometric accuracy of their exam.

Conclusion

The methods for answering the research questions include highly specialized procedures and cutting edge technology. To date, little has been done to test whether number correct scoring is a viable option in a ca-MST framework, this research study explores this by testing the overall resulting difference(s) in ability estimation when using number correct routing versus IRT routing.

CHAPTER IV

RESULTS

In this section the results from the ATA and simulation study are presented. The results from the ATA include how well the ATA engine performed in generating module information functions (MIFs) versus the targets (discussed in Chapter III). Then, results from the ca-MST simulations are presented.

Creation of the MIFs

This section discusses the results from the creation of the MIFs including how well the resulting MIFs from the ATA fit to the targets. A description of how the targets were generated can be found in Chapter III. Figure 4.1 displays an example of one of the MIF v Target plots generated. The black geometric figures are the actual MIFs while the colored lines are the target functions. The scale runs from 0 to the maximum information in the entire first configuration. If the plots fit well there would be very little difference between the colored line (target) and the geometric points (MIFs). In Figure 4.1 we see very little difference so we can assess that the target MIFs fit pretty well in this iteration. Figure 1 is by replication, so there are 12 modules per stage which means there are twelve module lines in the left side (stage 1) of Figure 4.1 and 4 module lines in each of the three MIFs on the right side of Figure 4.1.

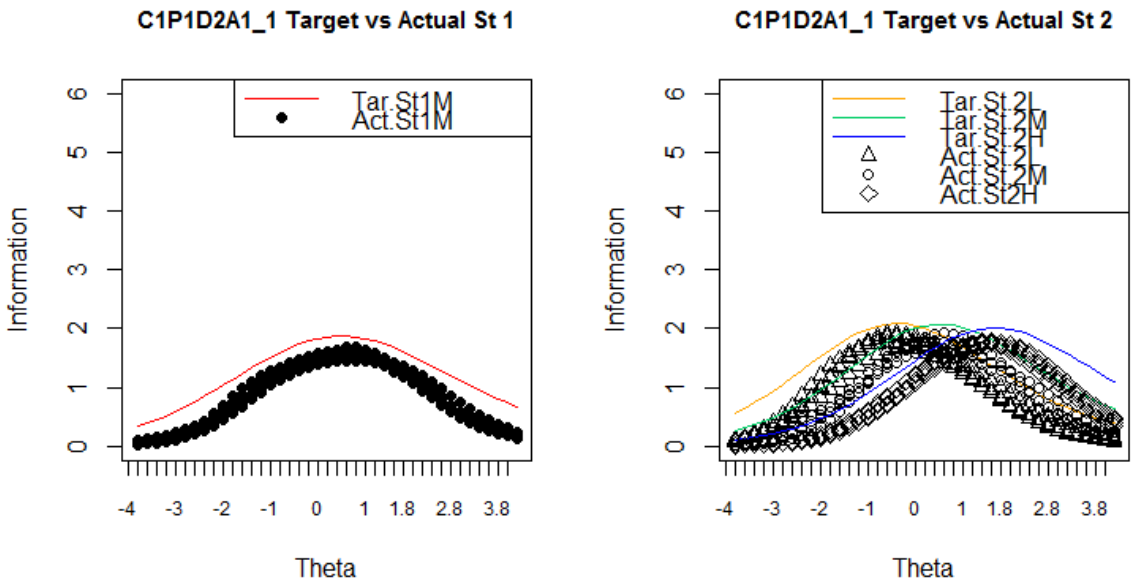

Figure 4.1. Example of 1-3 MIF design. This figure shows the fit of the MIFs to the MIF targets. The solid black geometric shapes are the actual MIFs used in the analysis, while the solid colored lines were the targets.

Since the study included ten replications per each of the 36 study conditions it's important to see if there are any particular outliers when it comes to the ATA process. As a result, the modules within each stage were averaged across condition for each of the ten replications. As shown in Figure 4.2, there is little variance between the conditions even when the number of modules within a stage is small (as is the case with the 3rd stage). This indicates that there is no real anomaly in the ATA for each of the 36 conditions. In this case there are 10 module lines (one for each replication averaged across the 12 modules). A different color was used for each MIF. As with Figure 4.1, closeness to the targets indicates good fit. In addition, if there's a MIF that diverges from the other MIFs that would indicate perhaps an aberrant replication of the specific condition.

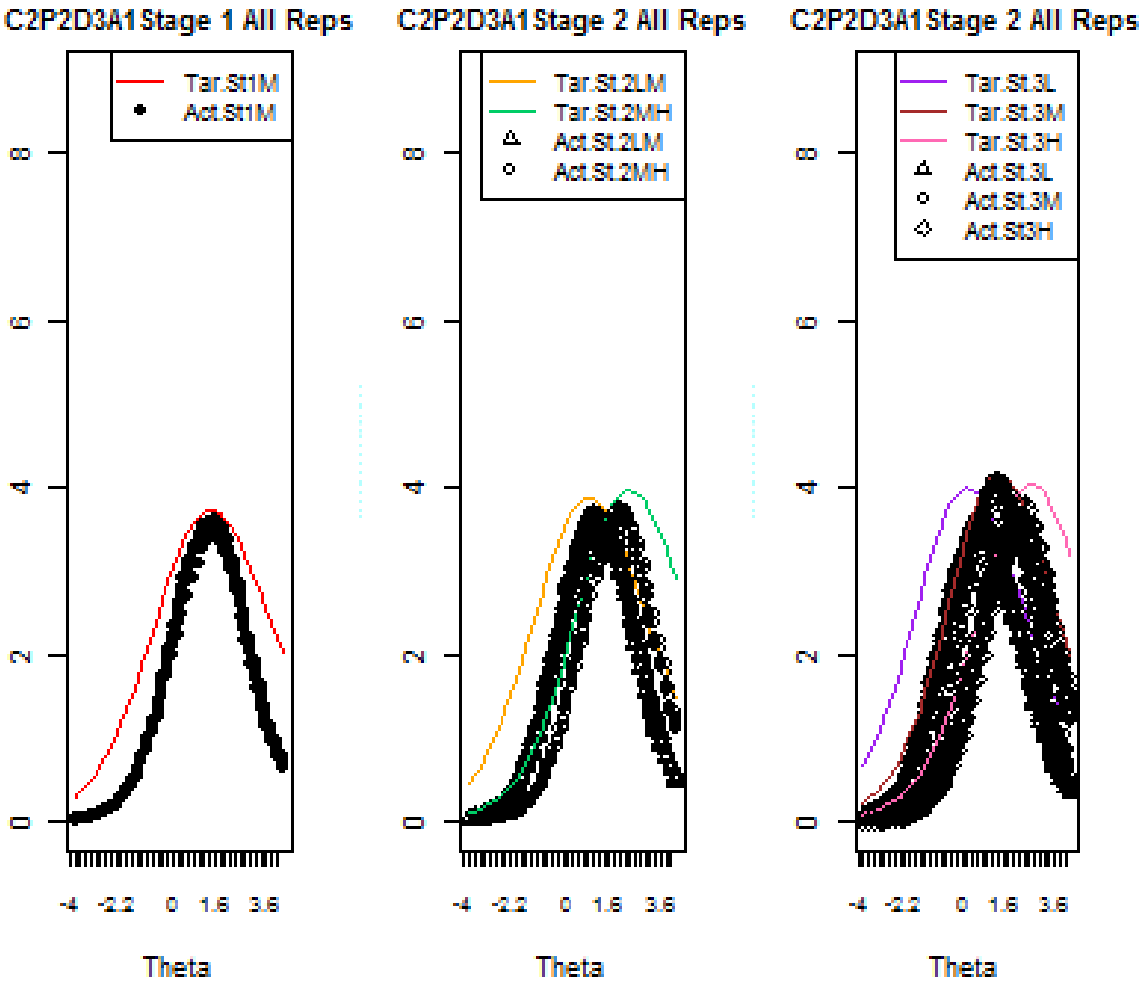

Figure 4.2. Example configuration TIF to target plot. This figure shows the fit of the MIFs to the MIF targets for each of the ten replications. The solid black geometric shapes are the actual MIFs used in the analysis, while the solid colored lines were the targets.

The overall MIF RMSEs of the actual MIFs to the target MIFs are displayed in Tables 4.1 to 4.3. Overall, the a-parameters seem to be the biggest difference in fit. The configurations with the a-parameter values with a mean of 0.6 had a better chance of fitting. Typically the fit was most compromised at tails of the distribution as seen in

Figures 4.1 and 4.2. This could be an artifact of the item parameter generation. The modules in the 1-3 configuration tend to fit a little bit better than the modules from either the 1-2-3 or the 1-2-3-4. The 1-2-3 and 1-2-3-4 fit equally well in terms of the RMSE.

These fit indices are summed over the entire range of the scale (-4 to 4), while there are great amounts of misfit towards the tails of the distribution, there is significantly less misfit where the peaks of the distribution are. Another point of interest might be the low standard deviations on the configuration means. While Figure 4.2 provides a picture, this table suggests that the target to actual relationship is similar for all 10 replications.

Table 4.1

RMSE of ATA-created MIFs vs. Target Functions for 1-3 Configurations

Config	mn-S1M	s-S1M	mn-S2L	s-S2L	mn-S2M	s-S2M	mn-S2H	s-S2H
C1	1.00	0.51	1.01	0.49	0.92	0.44	0.95	0.50
C1P1D1A1	0.44	0.00	0.45	0.02	0.44	0.01	0.44	0.01
C1P1D1A2	0.98	0.01	0.98	0.01	0.88	0.01	0.99	0.02
C1P1D2A1	0.43	0.01	0.44	0.01	0.41	0.01	0.39	0.01
C1P1D2A2	1.00	0.01	0.96	0.01	0.91	0.01	0.96	0.04
C1P1D3A1	0.39	0.01	0.41	0.01	0.37	0.01	0.33	0.02
C1P1D3A2	0.93	0.01	0.95	0.01	0.83	0.02	0.82	0.05
C1P2D1A1	0.86	0.02	0.93	0.06	0.86	0.04	0.85	0.04
C1P2D1A2	1.81	0.03	1.84	0.06	1.60	0.04	1.83	0.05
C1P2D2A1	0.82	0.03	0.86	0.05	0.82	0.04	0.77	0.05
C1P2D2A2	1.85	0.03	1.76	0.05	1.67	0.04	1.78	0.05
C1P2D3A1	0.74	0.01	0.77	0.05	0.67	0.02	0.66	0.04
C1P2D3A2	1.74	0.03	1.73	0.04	1.56	0.04	1.60	0.11

Table 4.2

RMSE of ATA-created MIFs vs. Target Functions for 1-2-3 Configurations

Config	mn-S1	s-S1	mn-S2LM	s-S2LM	mn-S2MH	s-S2MH	mn-S3L	sd-S3L	mn-S3M	s-S3M	mn-S3H	s-S3H
C2	1.08	0.60	1.07	0.59	1.04	0.61	1.11	0.66	1.02	0.56	1.08	0.67
C2P1D1A1	0.45	0.01	0.44	0.01	0.41	0.01	0.42	0.02	0.43	0.01	0.42	0.02
C2P1D1A2	1.00	0.01	0.93	0.01	0.91	0.01	0.96	0.01	0.91	0.01	0.93	0.02
C2P1D2A1	0.42	0.00	0.41	0.01	0.40	0.01	0.41	0.01	0.40	0.01	0.39	0.01
C2P1D2A2	0.95	0.01	0.91	0.01	0.94	0.02	0.92	0.01	0.92	0.01	0.94	0.03
C2P1D3A1	0.38	0.01	0.40	0.01	0.34	0.01	0.42	0.02	0.37	0.01	0.34	0.01
C2P1D3A2	0.92	0.01	0.89	0.01	0.87	0.02	0.90	0.01	0.84	0.02	0.86	0.05
C2P2D1A1	0.93	0.03	0.97	0.04	0.87	0.02	0.89	0.03	0.90	0.04	0.92	0.05
C2P2D1A2	2.06	0.03	2.20	0.05	1.87	0.03	2.54	0.05	1.88	0.02	2.08	0.07
C2P2D2A1	0.93	0.02	0.91	0.04	0.85	0.06	0.90	0.05	0.88	0.05	0.91	0.07
C2P2D2A2	2.00	0.04	1.88	0.04	2.07	0.09	1.90	0.06	1.99	0.06	2.08	0.09
C2P2D3A1	0.84	0.02	0.91	0.03	0.75	0.04	1.03	0.07	0.87	0.05	0.75	0.06
C2P2D3A2	2.07	0.06	1.94	0.05	2.15	0.07	2.06	0.09	1.89	0.08	2.33	0.07

Table 4.3

RMSE of ATA-created MIFs vs. Target Functions for 1-2-3-4 Configurations

Config	mn-S1	s-S1	mn-S2LM	s-S2LM	mn-S2MH	s-S2MH	mn-S3L	sd-S3L	mn-S3M	s-S3M
C3	1.06	0.57	1.08	0.58	1.08	0.63	1.13	0.63	0.98	0.51
C3P1D1A1	0.45	0.01	0.45	0.01	0.44	0.01	0.45	0.02	0.42	0.01
C3P1D1A2	0.95	0.01	0.94	0.01	0.93	0.02	0.92	0.01	0.87	0.01
C3P1D2A1	0.43	0.01	0.43	0.01	0.43	0.01	0.46	0.01	0.41	0.01
C3P1D2A2	0.98	0.01	0.96	0.01	0.92	0.02	0.99	0.01	0.83	0.02
C3P1D3A1	0.37	0.00	0.42	0.01	0.36	0.01	0.42	0.01	0.38	0.01

Table 4.3

(Cont.)

Config	mn-S1	s-S1	mn-S2LM	s-S2LM	mn-S2MH	s-S2MH	mn-S3L	sd-S3L	mn-S3M	s-S3M
C3P1D3A2	0.89	0.01	0.87	0.01	0.90	0.03	0.90	0.01	0.85	0.02
C3P2D1A1	0.93	0.02	0.95	0.04	0.92	0.02	0.98	0.07	0.86	0.04
C3P2D1A2	1.94	0.02	2.24	0.03	1.96	0.03	2.31	0.11	1.75	0.02
C3P2D2A1	0.97	0.02	0.92	0.04	0.97	0.04	1.03	0.05	0.90	0.07
C3P2D2A2	2.01	0.05	1.97	0.04	1.98	0.07	2.09	0.05	1.69	0.08
C3P2D3A1	0.79	0.02	0.97	0.03	0.78	0.04	0.98	0.06	0.82	0.04
C3P2D3A2	1.96	0.05	1.78	0.03	2.34	0.11	1.98	0.05	1.93	0.05

Config	mn-S3H	s-S3H	mn-S4L	s-S4L	mn-S4LM	s-S4LM	mn-S4MH	s-S4MH	mn-S4H	s-S4H
C3	1.08	0.67	1.13	0.66	1.01	0.55	1.01	0.56	1.08	0.65
C3P1D1A1	0.42	0.01	0.47	0.03	0.43	0.01	0.43	0.01	0.45	0.01
C3P1D1A2	0.95	0.02	0.91	0.01	0.93	0.01	0.91	0.02	0.88	0.02
C3P1D2A1	0.42	0.02	0.46	0.02	0.43	0.01	0.42	0.01	0.40	0.02
C3P1D2A2	0.94	0.01	0.92	0.01	0.86	0.02	0.87	0.01	0.91	0.03
C3P1D3A1	0.34	0.01	0.42	0.01	0.37	0.01	0.35	0.01	0.34	0.02
C3P1D3A2	0.81	0.03	0.88	0.01	0.85	0.01	0.84	0.02	0.90	0.05
C3P2D1A1	0.89	0.04	0.97	0.06	0.89	0.05	0.88	0.03	1.03	0.08
C3P2D1A2	2.12	0.05	2.68	0.08	2.03	0.04	1.88	0.04	1.90	0.05
C3P2D2A1	0.93	0.04	1.05	0.05	0.90	0.04	0.93	0.05	0.94	0.05
C3P2D2A2	2.20	0.13	1.90	0.06	1.80	0.05	1.88	0.07	2.06	0.11
C3P2D3A1	0.72	0.03	0.97	0.04	0.85	0.05	0.76	0.06	0.77	0.05
C3P2D3A2	2.17	0.05	1.91	0.03	1.83	0.06	1.94	0.13	2.41	0.08

Overall, the TIFs fit for most of the distribution. The misfit tended to occur where the mean discrimination in the item pool was high and when the item pool difficulty was off-center. Most of the misfit tended to be in the tails. The shape of the TIFs tended to match the targets, however, the amount of information tended to be slightly depressed in the tails of the distribution.

Differences in Routing

Because of the way that the study is set up, the only way that a candidate's ability estimate differs is if the route changes as a function of the routing/scoring method. Figure 4.3 illustrates the effect of the item pool difficulty on routing for the 1-3 design. Overall, the exposure control method seems to be working as it's supposed to routing roughly 33% of examinees to one of the three routes taken. In addition, moderate to moderate ("M-M") route appears to be equally used in all four of the routing methods. One thing to note is that both the moderate to low route ("M-L") and the moderate to high route ("M-H") both have considerable variation to the amounts of students routed to each module.

Figure 4.4 completes the story set up by Figure 4.3.When the item pool is centered at either -1.0 or 1.0 and examinees are routed via maximum information routing, the examinees are routed primarily to the highest or lowest module more often than when the average difficulty of the item pool is centered at 0.0. When the item pool is centered at 0.0, the proportions of candidates sent to any of the 2nd stage modules is roughly equal; however, when the difficulty is at -1.0 or 1.0, the population is either sent to the highest module (as in the case of -1.0 average item difficulty) or to the lowest module (as

in the case of the 1.0 average item difficulty) disproportionately. This effect is not seen in exposure control routing because it control the proportions of examinees that are sent to each route.

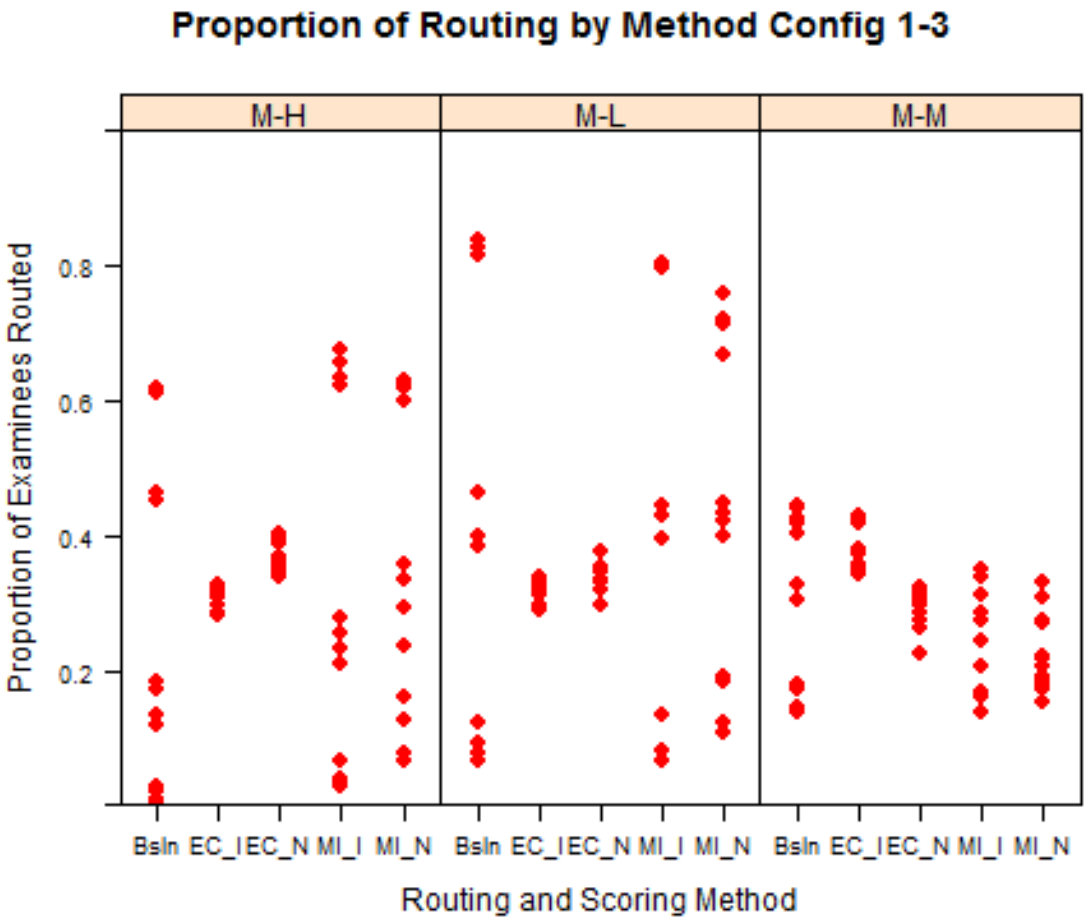

Figure 4.3. Proportion of routing by method. This figure shows the 5 methods by proportion of examinees routed to each of the routes in a 1-3 design. The names of the four scoring/routing methods have been abbreviated: Bsln = Baseline, EC_I = Exposure Control routing + IRT scoring, EC_N = Exposure Control routing and Number Correct scoring, MI_I = Maximum Information routing and IRT scoring and MI_N is Maximum Information routing and Number Correct scoring.

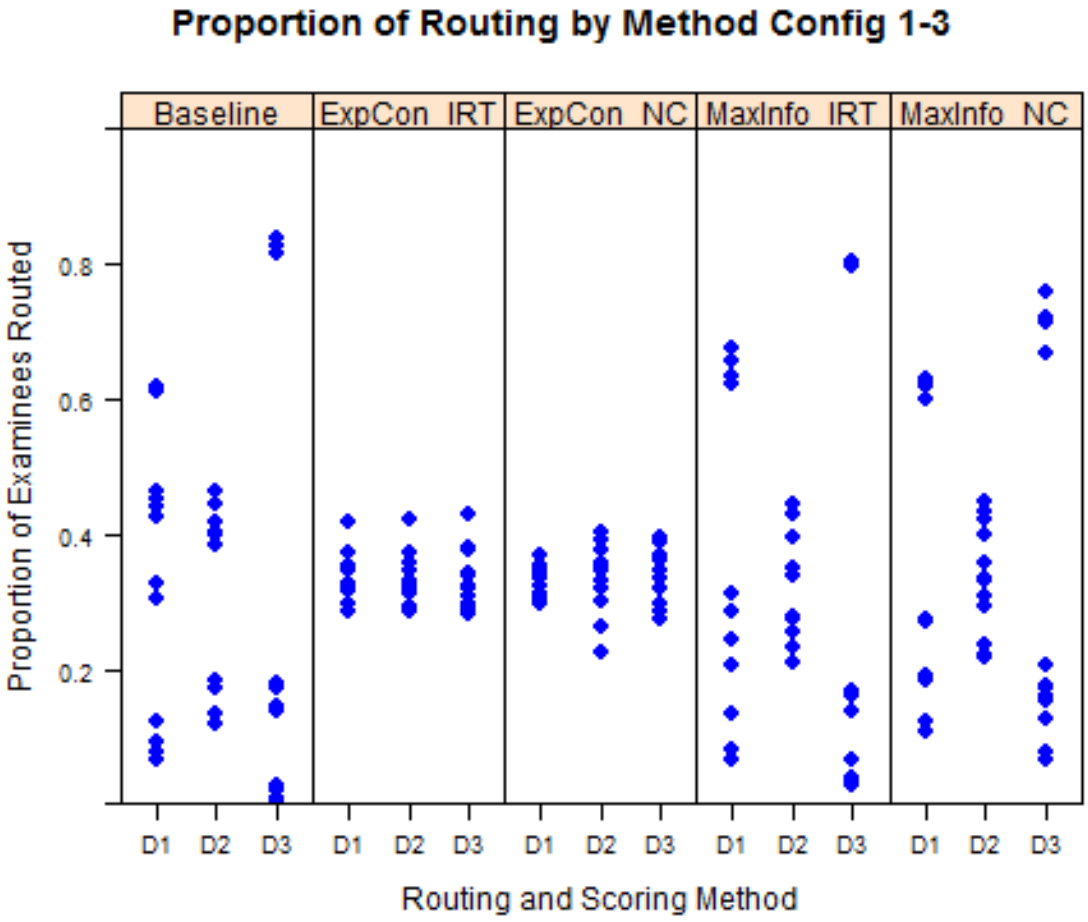

Figure 4.4. Proportion of routing by configuration. This shows the differences in routing by mean difficulty by routing and scoring method. D1 = -1.00, D2 = 0.00 and D3 = 1.00.

The 1-2-3 has a similar pattern to it seen in Figure 4.5. The middle two routes are very tight in terms of their proportions routed in each of the four methods; however, in the two tail patterns for the maximum information routing methods, there are cases where the proportions of people taking that particular route is over .80.

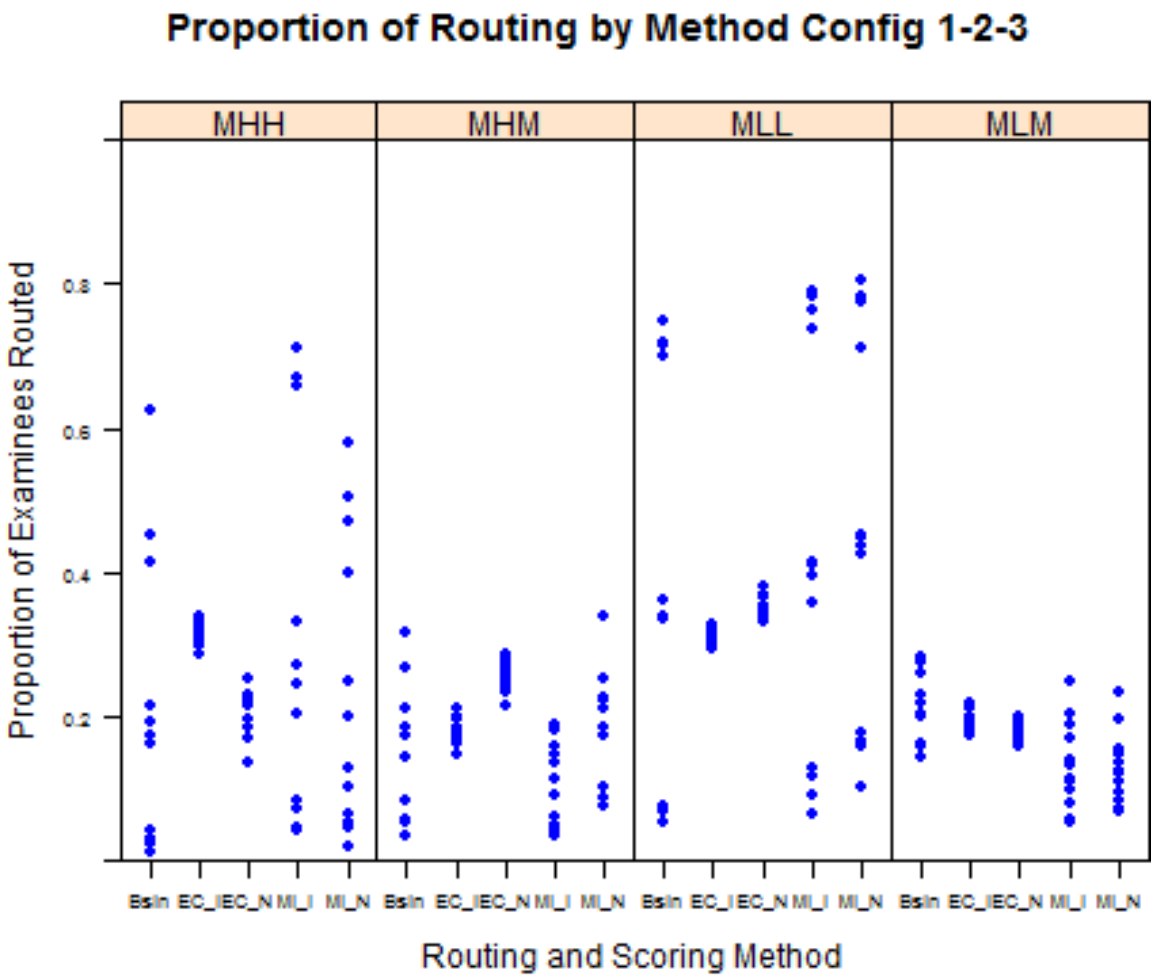

Figure 4.5. Proportion of routing scores by method in the 1-2-3 configuration. This figure is similar to Figure 4.3 using the 1-2-3 configuration. Bsln= Baseline, EC_I = Exposure Control routing + IRT scoring, EC_N = Exposure Control routing and Number Correct scoring, MI_I = Maximum Information routing and IRT scoring and MI_N is Maximum Information routing and Number Correct scoring.

Overall, we see the same pattern in terms of routing in each of the four methods. The maximum information methods tend to route people differently than the exposure

control methods particularly when the mean difficulty in the item pool is off center (b = -1.00 or b = 1.00).

Residuals and Recovery

Ultimately the overall goal of any testing method is to provide the most accurate information possible about the examinee's ability level. No matter the choice in ca-MST configuration, routing method or scoring method, the ca-MST should be constructed to provide precise information about the examinee's ability. This section of the dissertation examines the overall effects of the routing and scoring decisions and configuration decisions on the recovery of examinee ability. First, the overall graphs will be shown for each of the conditions in the study. Next, tables and figures will be displayed with the collapsed information across replications. Each of these will be broken up by each of the four routing/scoring methods. Each plot will contain the results of the "baseline" which is an unrestricted ca-MST that essentially behaves like a maximum information CAT routing the examinee to the items that are most appropriate for the examinee regardless of the route.

Each of the below figures has the scale set a little higher than the max RMSE and runs from 0.0 to the max of the condition plus a little bit. This allows all of the points to be seen on the plot and allows also examining of the patterns in the data. The bias plots run from -0.10 to 0.15. The colors are consistent across all of the plots with a color per method: red represents the baseline statistic, blue indicates the maximum information routing and IRT scoring method, green indicates the maximum information routing and number correct scoring method, purple indicates the exposure control routing and IRT

scoring method, and pink represents the exposure control routing and number correct scoring method.

Item Pool Characteristics

Each of the item pool characteristics was tested to see if the levels in them caused changes in the overall ability recovery. The item pool characteristics in this section are item pool difficulty and average discrimination.

Three item pool difficulties were chosen to observe the effect of different item pool difficulties on the configuration decisions. The three levels of item pool difficulty were low difficulty pool (D1 = -1.00), a moderate difficulty pool (D2 = 0.00), and a high difficulty pool (D3 = 1.00). Figure 4.6 displays the bias by mean difficulty of the item pool. There is little difference in bias across the three levels of difficulty. One of the -1.00 pools ended up having some slight bias in the positive direction for the exposure control metrics; however, this effect was not consistent across replications within or across conditions. On average, D1 and D2 contained a little more variation than D3 did.

Figure 4.7 displays the RMSE for the average pool difficulty and routing and scoring method. Overall, there were no major differences in either of the difficulty methods. While this result may not be surprising, it's good to note that the changes in pool difficulty do not affect the decision made in routing method. D2 tended to be slightly lower however still overlaps with the majority of the two off-center difficulties. In addition, D3 tended to produce the highest RMSE results out of any of the three difficulties. D2 tends to dip slightly below the other two difficulty measures as is illustrated by Figure 4.7, one possible explanation could be that D2 matches the

generated ability distribution and thus tends to recover ability just a little bit better. D1 tends to range from 0.25 to 0.55 and D3 tends to range from 0.25 to above 0.60, whereas D2 tends to range from 0.20 to 0.50. While the RMSE tends not to be much lower at its lowest, the ceiling is slightly lower.

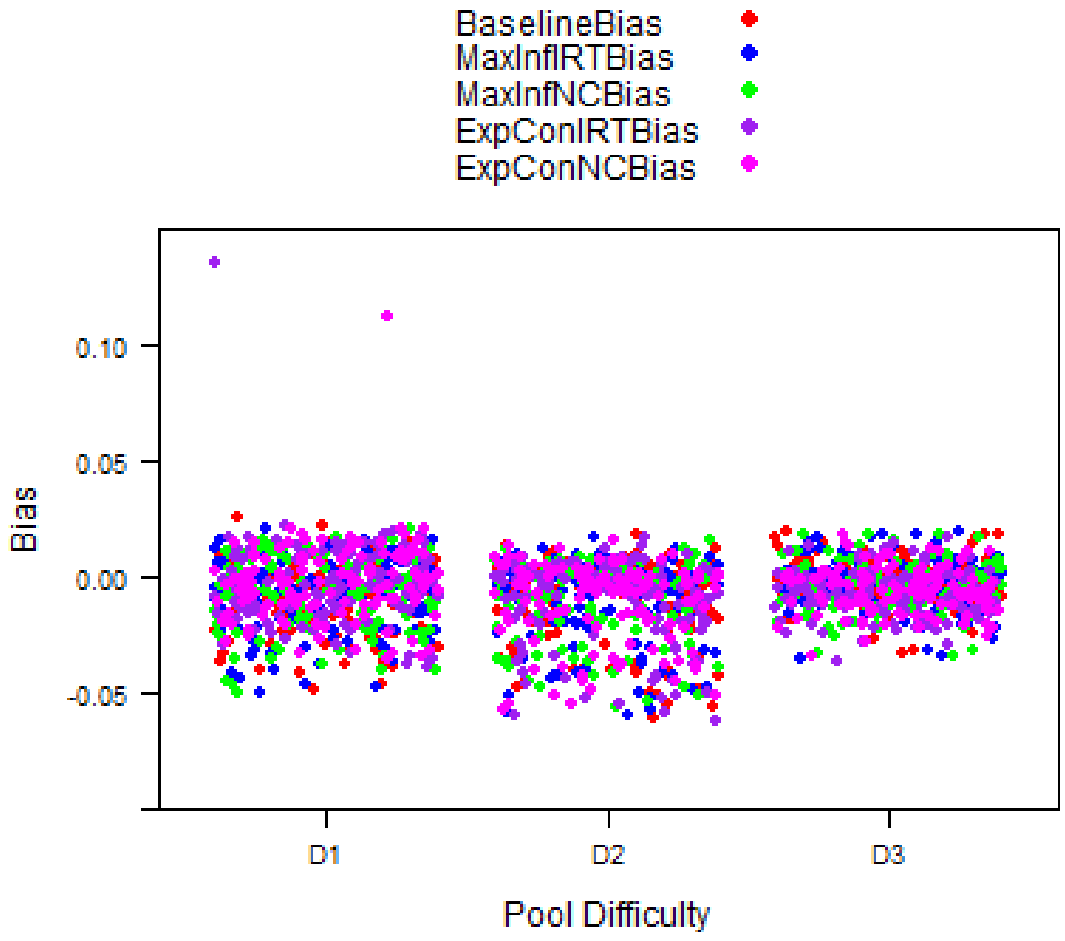

Figure 4.6. Bias by method and pool difficulty. MaxInfIRT = Maximum Information routing with IRT scoring, MaxInfNC = Maximum Information Routing with Number-correct scoring, ExpConIRT = Exposure Control routing with IRT scoring and ExpConNC = Exposure Control Routing with Number-correct scoring.

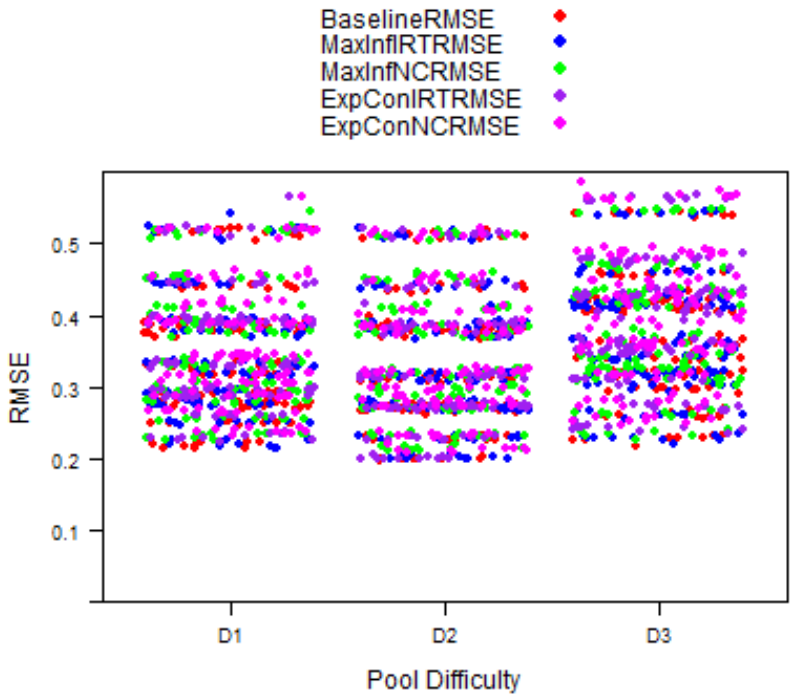

Figure 4.7. RMSE by pool difficulty and routing/scoring method. MaxInfIRT = Maximum Information routing with IRT scoring, MaxInfNC = Maximum Information Routing with Number-correct scoring, ExpConIRT = Exposure Control routing with IRT scoring and ExpConNC = Exposure Control Routing with Number-correct scoring.

The other item pool characteristic included in this study is the average discrimination for the item pool. The two levels of average item discrimination chosen for this study were low average item discrimination (A1 = 0.60) and high average item discrimination (A2 = 1.00). The two levels of average item discrimination in the pool were compared by bias and RMSE. Figures 4.8 and 4.9 illustrate the differences in these two statistics.

Figure 4.8 displays the bias by both method and average discrimination. On average, both levels of discrimination appear to be around 0.00 average bias. It is important to note; however, that when the average discrimination is high (a = 1.00) that in this study, some of the conditions ended up under estimating ability.

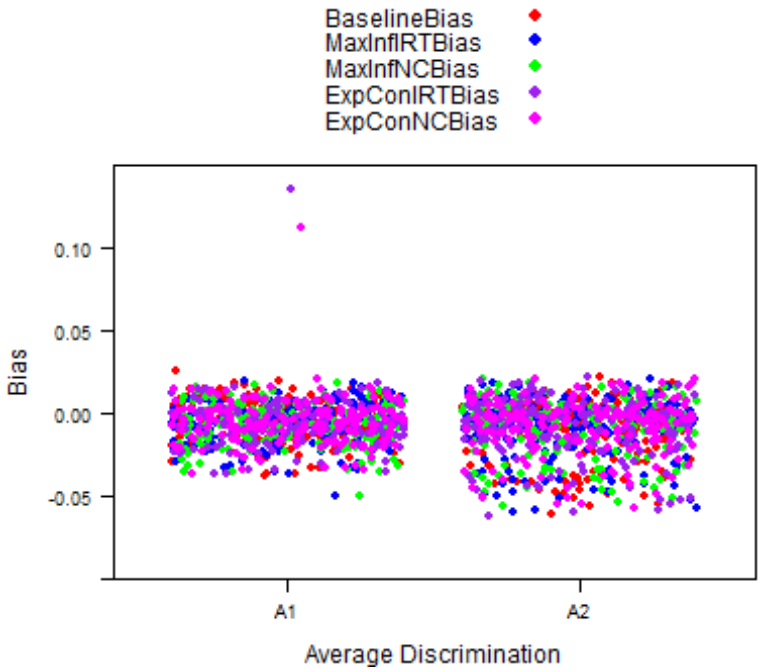

Figure 4.8. Bias by method and average discrimination. MaxInfIRT = Maximum Information routing with IRT scoring, MaxInfNC = Maximum Information Routing with Number-correct scoring, ExpConIRT = Exposure Control routing with IRT scoring and ExpConNC= Exposure Control Routing with Number-correct scoring.

Figure 4.9 displays the RMSE for each of the four methods by the average discrimination. On average the RMSE for the higher average discrimination is lower. This finding is the result of items that are higher quality in terms of the amount of information each item if the items are of higher quality, information will be higher which

will result in a drop of the standard error or more precise estimation which will result in lower RMSE values. Values of RMSE range from just below 0.30 to roughly 0.60 for A1 and range from roughly 0.20 to just under 0.50 for A2. Figure 4.9 shows the A2 RMSE quite a bit below the RMSE for the A1 discrimination condition.

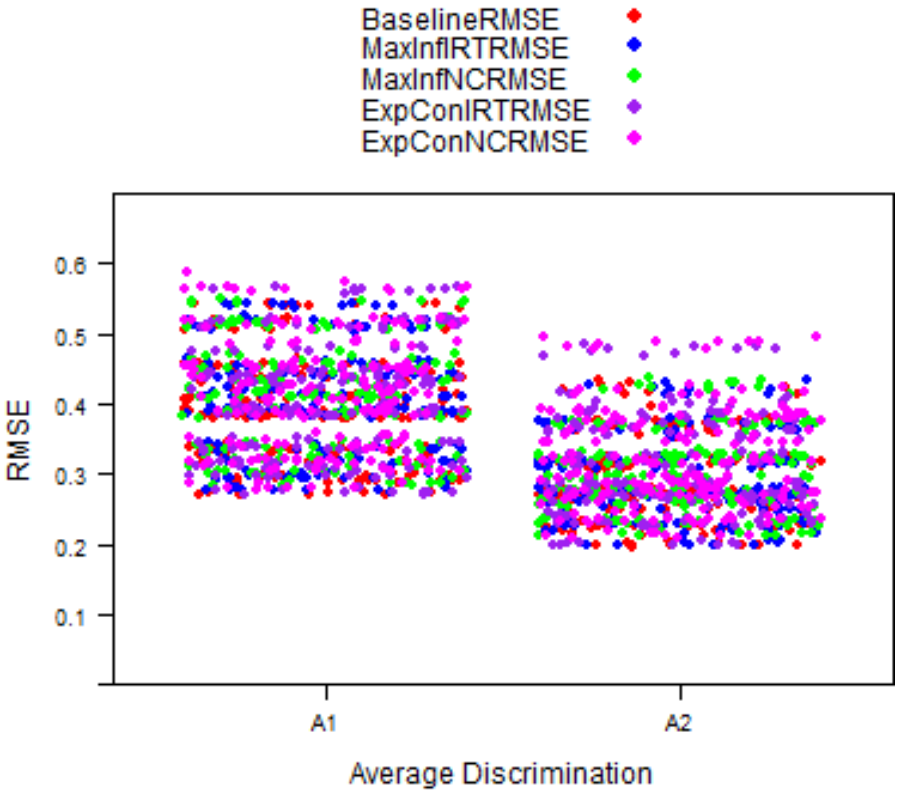

Figure 4.9. RMSE by method and average discrimination. MaxInfIRT = Maximum Information routing with IRT scoring, MaxInfNC = Maximum Information Routing with Number-correct scoring, ExpConIRT = Exposure Control routing with IRT scoring and ExpConNC= Exposure Control Routing with Number-correct scoring.

Overall, there were only a few differences in the item pool characteristics. First, the bias for the average discrimination tended to be slightly under-estimated when the average discrimination was a little higher. Second, when the average discrimination of the item pool tended to be higher, the RMSE tended to be lower.

Ca-MST Configuration Characteristics

Similarly to the item pool characteristics, each of the ca-MST configuration characteristics was tested to see if the levels in them caused changes in the overall ability recovery. There were two ca-MST configuration characteristics in this study: three levels of ca-MST designs/configurations and two levels of module length/item pool size. Each of these conditions will be examined in terms of bias and RMSE and plots/tables will be displayed to describe the differences in ability recovery for each of these two variables.

Figure 4.10 displays the bias results for the two different module lengths (m = 10 items and m = 20 items). The module length tended to result in about 0 bias for each of the two module lengths; however, when the test was longer, the bias tended to be slightly more variable, whereas the lower sized modules tended to be tightly bound around 0.00.

Figure 4.11 illustrates the same results but instead of bias the RMSE results are shown. The results here are notable if unexpected. The longer the module length is the lower the RMSE is. More items mean lower errors and lower errors mean better, more precise results. The lower module length (m = 10 items) ranges from just below 0.30 to roughly 0.60. The higher module length (m = 20 items) ranged from roughly 0.20 to 0.40 with one case being exceptionally high at roughly 0.50.

Lastly, the three ca-MST designs were compared in terms of bias and RMSE. The three configurations in this study were the 1-3 (C1), 1-2-3 (C2), 1-2-3-4 (C3). These are three of a nearly infinite pool of ca-MST designs that could be chosen. These pools were chosen because they vary the number of stages and levels of difficulty within a stage.

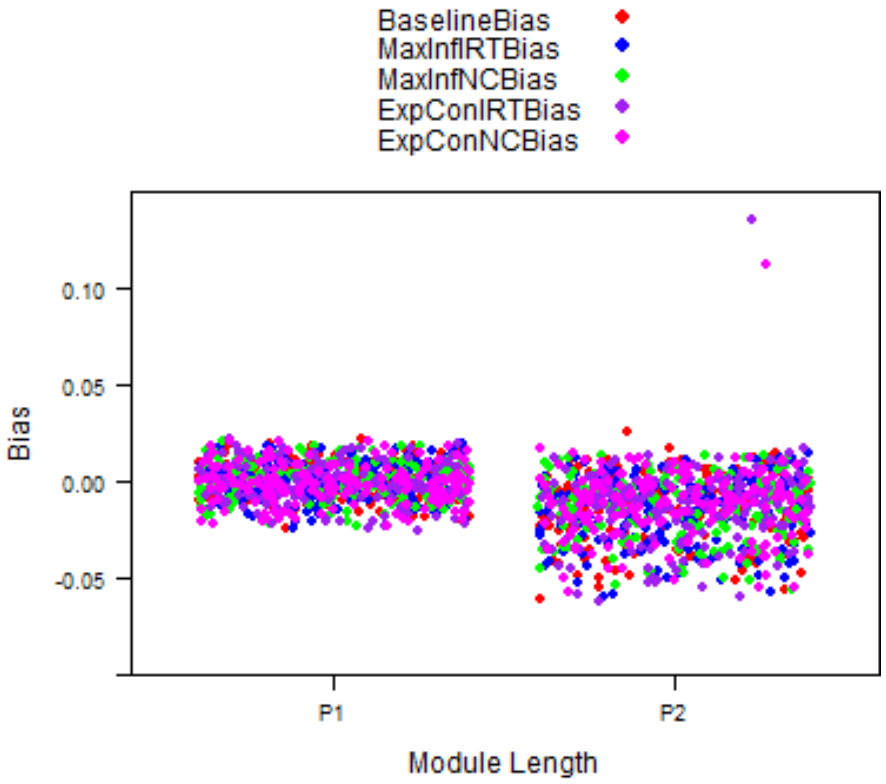

Figure 4.10. Bias by method and module length. MaxInfIRT = Maximum Information routing with IRT scoring, MaxInfNC = Maximum Information Routing with Number-correct scoring, ExpConIRT = Exposure Control routing with IRT scoring and ExpConNC= Exposure Control Routing with Number-correct scoring.

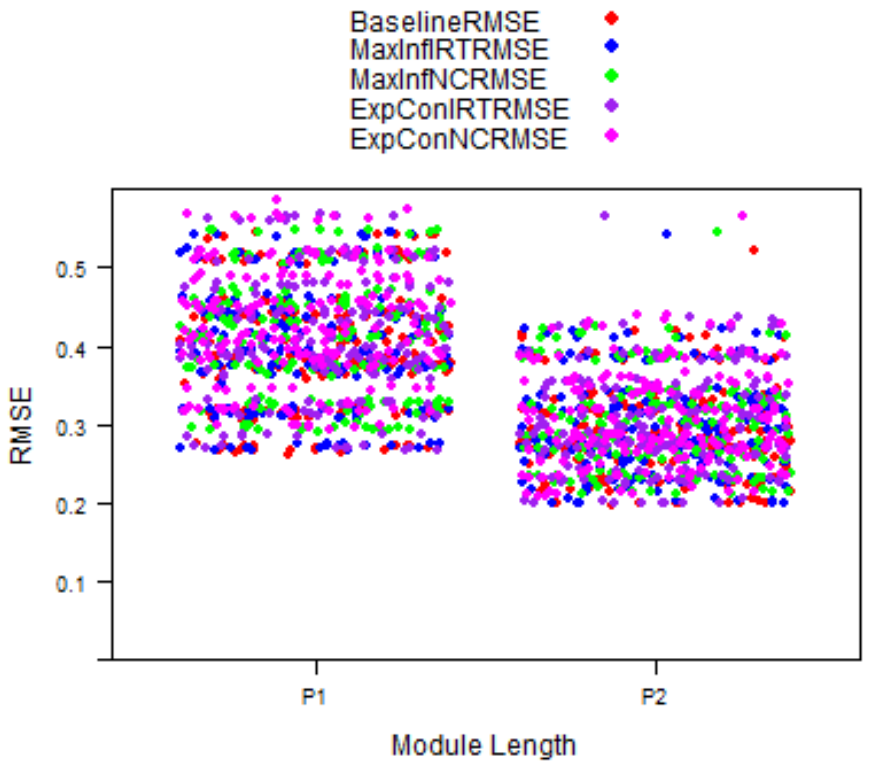

Figure 4.11. RMSE by method and module length. MaxInfIRT = Maximum Information routing with IRT scoring, MaxInfNC = Maximum Information Routing with Number-correct scoring, ExpConIRT = Exposure Control routing with IRT scoring and ExpConNC= Exposure Control Routing with Number-correct scoring.

Figure 4.12 depicts the average bias by each of the three configurations chosen for this study. Nothing here really sticks out. Configuration C1 might have slightly less variation in the bias statistic; however, all three configurations are roughly centered at 0.00 with very little variation.

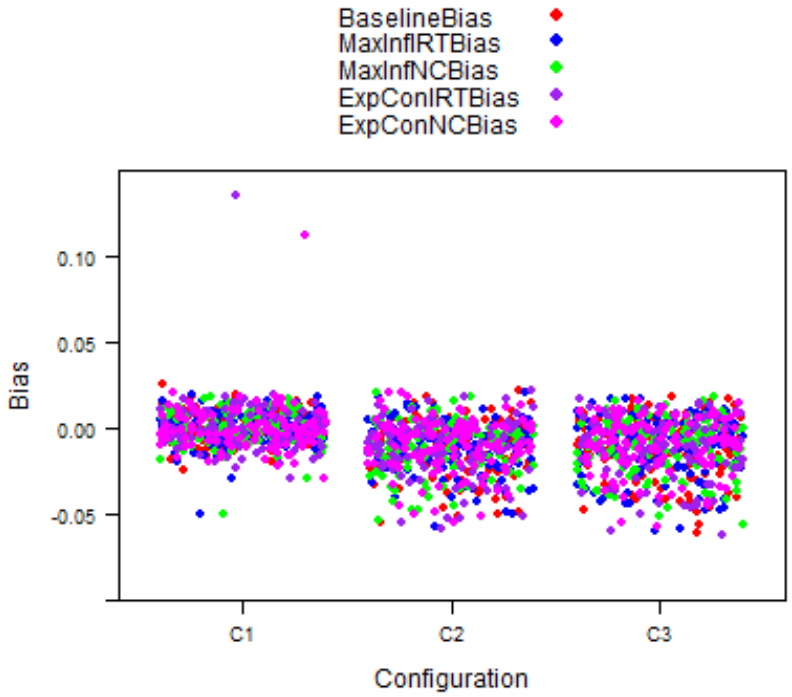

Figure 4.12. Bias by configuration and method. MaxInfIRT = Maximum Information routing with IRT scoring, MaxInfNC = Maximum Information Routing with Number-correct scoring, ExpConIRT = Exposure Control routing with IRT scoring and ExpConNC= Exposure Control Routing with Number-correct scoring.

Figure 4.13 highlights the difference in RMSE by configuration. Similar to the module length there is a decrease in terms of the number of stages. The more items an examinee sees the better the model will be able to recover ability. There is a rather steep drop in RMSE from C1 (1-3) to C2 (1-2-3) the extra module produces a rather stark drop in terms of recovery. The drop from C2 to C3 (1-2-3-4) is not as distinct.

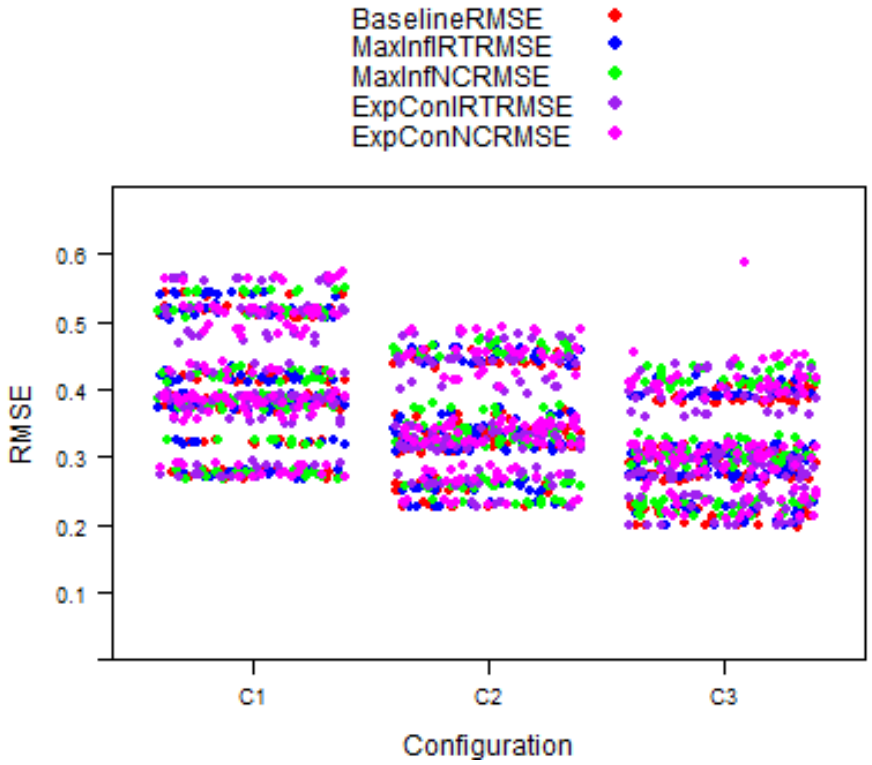

Figure 4.13. RMSE by configuration and method. MaxInfIRT = Maximum Information routing with IRT scoring, MaxInfNC = Maximum Information Routing with Number-correct scoring, ExpConIRT = Exposure Control routing with IRT scoring and ExpConNC = Exposure Control Routing with Number-correct scoring.

Overall, the effects of the ca-MST configurations tend to be where more items seen leads to lower amounts of RMSE. There are however, diminishing returns when it comes to adding items to the configuration. There are multiple ways to add items to the ca-MST either by increasing the size of the modules or the number of stages in the ca-MST design.

Routing and Scoring Methods

This section breaks the four routing and scoring sections down and summarizes the bias and RMSE results. Plots will be displayed to summarize both the routing and scoring and an overall table of all 36 conditions will be displayed. Figure 4.14 summarizes the bias plot for the two routing methods + the baseline condition which was a maximum information CAT with unrestricted routing. Maximum Information routing tended to result in slightly less bias than either the Exposure Control or Baseline methods. All three methods of routing tended to be centered around zero and ranged between -.05 and .02. The outliers found in EC are from one iteration that seems to have been an aberrant case.

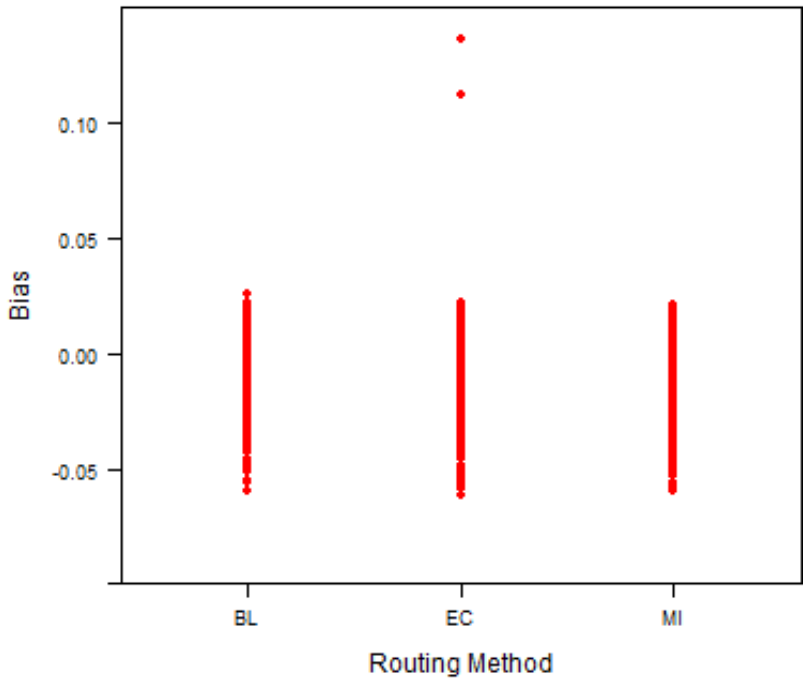

Figure 4.14. Bias by routing method. BL is the baseline, EC is exposure control routing and MI is maximum information routing.

Similarly, the RMSE for the routing method tells a similar story. Maximum information routing and the baseline tend to be a little lower in terms of RMSE. All three methods tended to have their minimum RMSE values around 0.20. Baseline tends to have a little less variance in RMSE than either the max information condition or the exposure control condition. Figure 4.15 displays the RMSE values for all 360 replications.

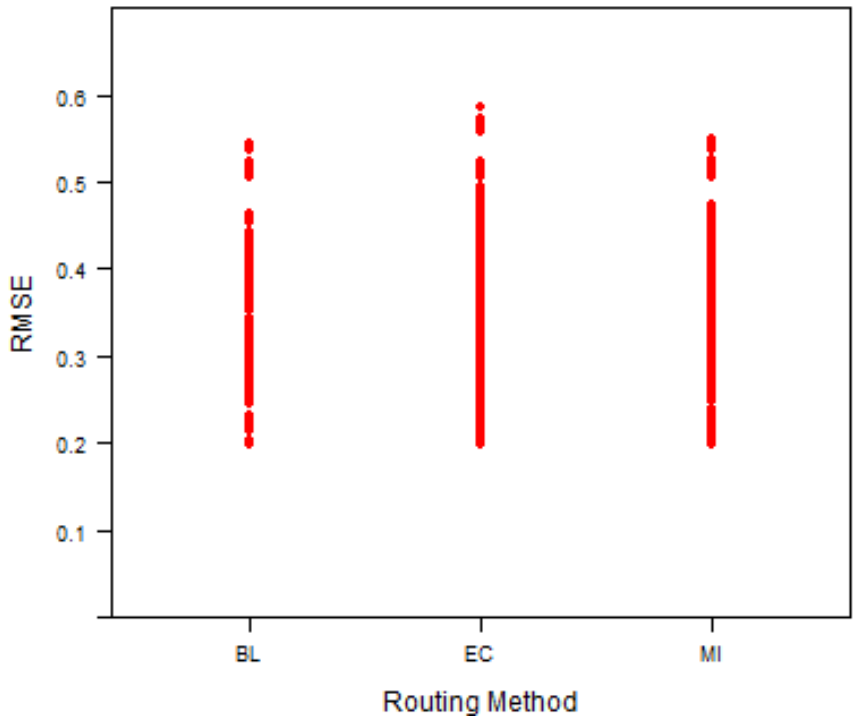

Figure 4.15. RMSE by routing method. BL is the baseline condition, EC is exposure control routing and MI is maximum information routing.

The two scoring methods in this paper were the IRT (EAP) approach versus the Number-Correct scoring method. In addition, the baseline method, although it uses IRT to score, was considered a category of its own. Figure 4.16 displays the bias statistics for

the three scoring methods. The IRT and NC methods are extremely similar to each other in terms of the overall distribution of bias. Figure 4.17 represents the distribution of RMSE statistics for the number correct scoring method versus the IRT method. Overall, the baseline CAT does slightly better than the IRT and NC methods; however, the difference is not great. In addition, the IRT method does slightly better than the NC method.

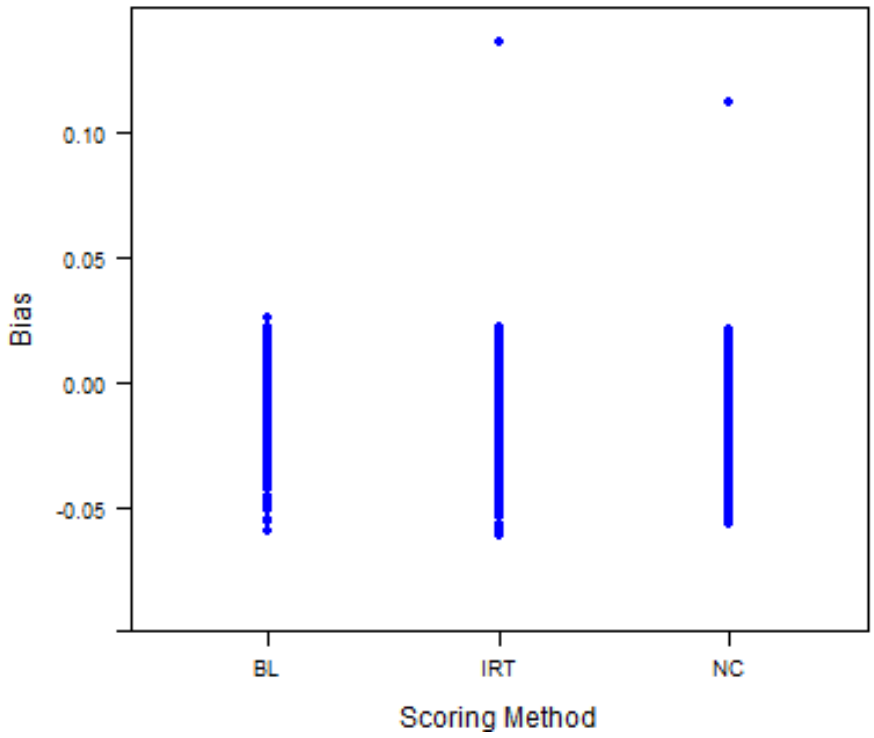

Figure 4.16. Bias by scoring method. BL is baseline scoring, IRT is IRT scoring and NC is number correct scoring.

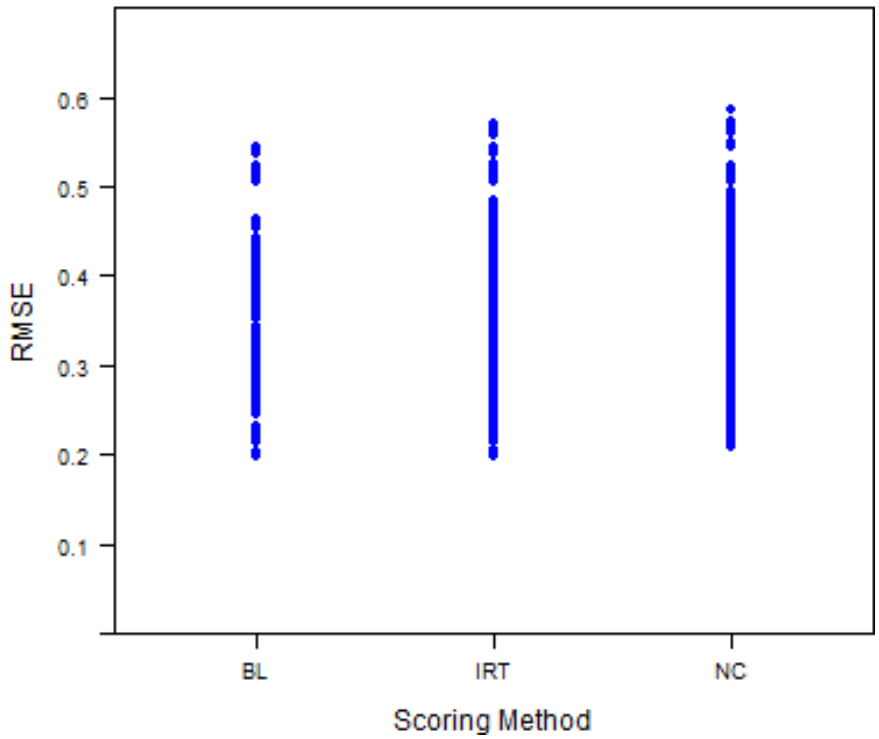

Figure 4.17. RMSE by scoring method. BL is baseline scoring, IRT is IRT scoring and NC is number correct scoring.

Figure 4.18 is the interaction of the two routing and scoring methods compared with the baseline in terms of RMSE. The maximum information IRT method performs nearly identically to the Baseline CAT method. The maximum information- number correct method has a slightly higher median value; however, is just a hair worse than the corresponding IRT method. In addition, the maximum information routing method tended to do a bit better than the exposure control methods. There does not really appear to be an interaction effect between the two.

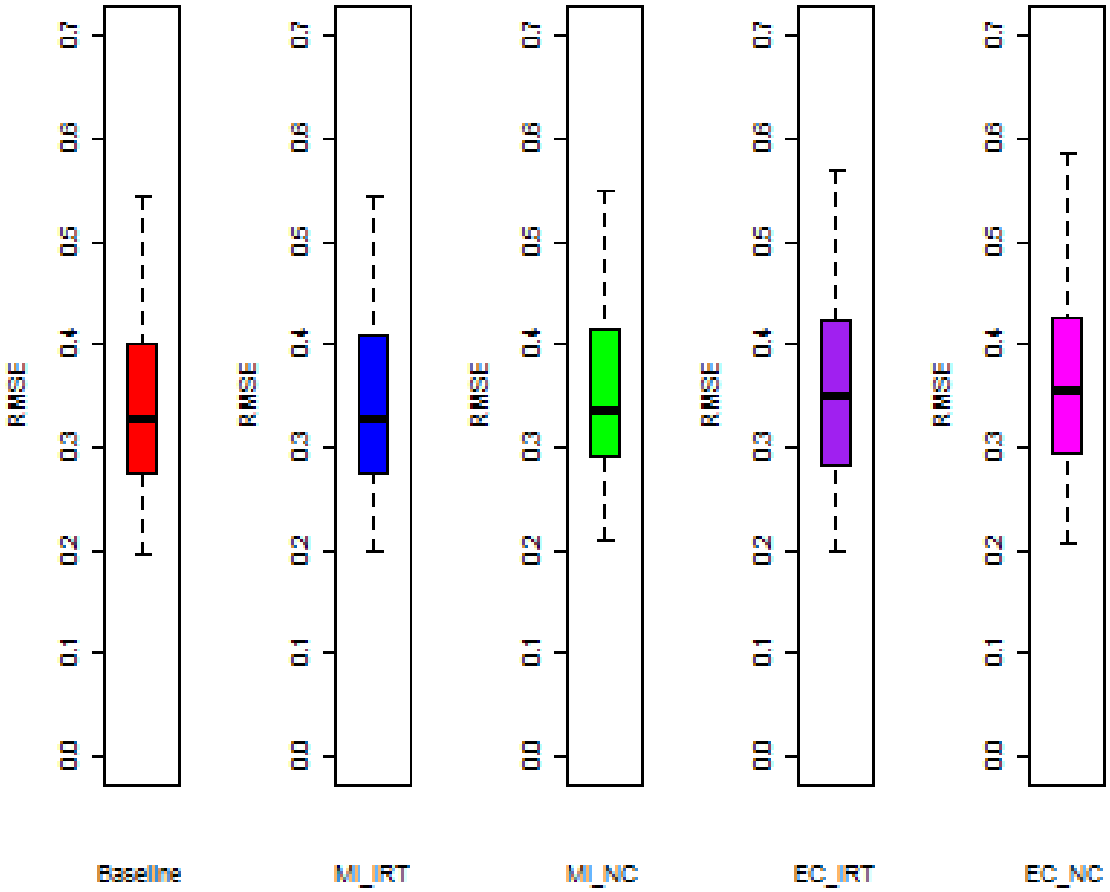

Figure 4.18. Boxplots of the interaction between the routing and scoring methods. Baseline = Baseline, EC_I = Exposure Control routing + IRT scoring, EC_N = Exposure Control routing and Number Correct scoring, MI_I = Maximum Information routing and IRT scoring and MI_N is Maximum Information routing and Number Correct scoring.

As seen in some of the plots of bias by routing and scoring method, the bias for some of the conditions was negative. Figure 4.19 examines the bias by condition in the study. Most of the boxes are right around 0.00, however, there are a couple conditions that are causing most of the issues with the bias statistic. When you look at those same

conditions in terms of RMSE however, as in Figure 4.20, the effect of the bias washes out as the RMSE tends to be lower relative to other similar conditions.

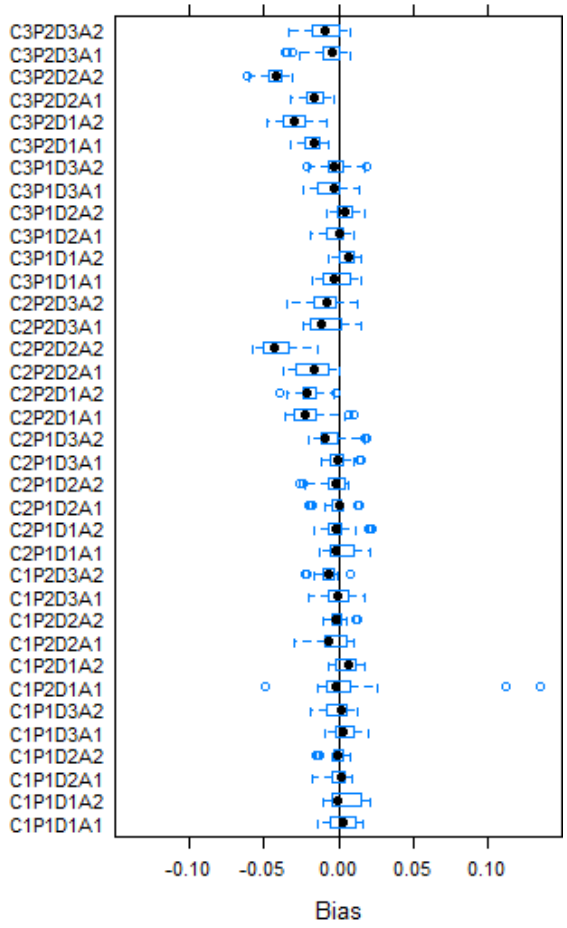

Figure 4.19. Boxplots of the average bias by each of the ten replications in the study for each condition. The configurations run from C1 (1-3) to C3 (1-2-3-4). The P is the module size ranging from P1 = 10 items per module to P2 = 20 items per module. D represents the average pool difficulty with values of D1 (-1.00), D2 (0.00), and D3 (1.00). Lastly, A is the average discrimination in the pool with values of A1 = 0.6 and A2 = 1.00.

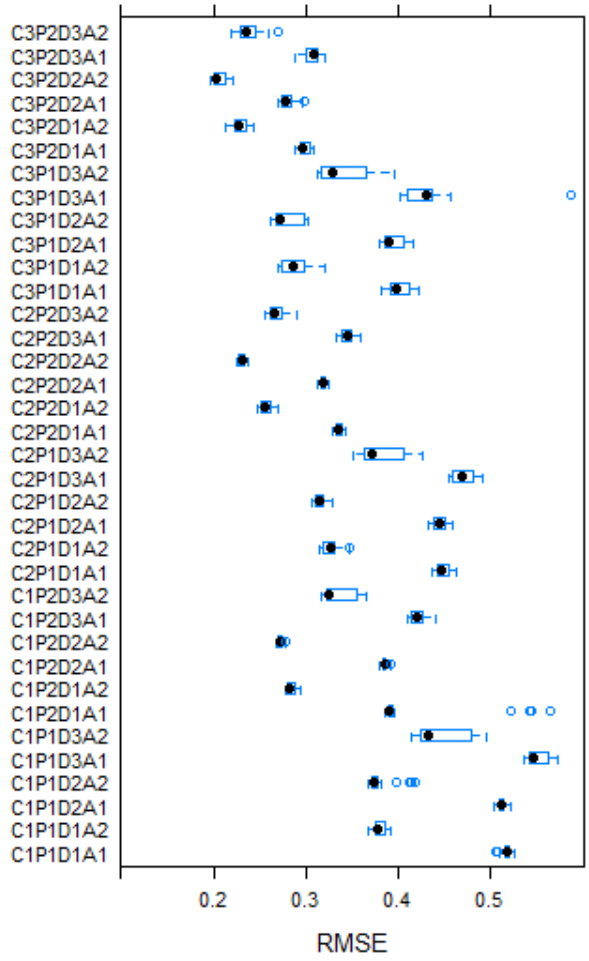

Figure 4.20. RMSE by condition. The configurations run from C1 (1-3) to C3 (1-2-3-4). The P is the module size ranging from P1 = 10 items per module to P2 = 20 items per module. D represents the average pool difficulty with values of D1 (-1.00), D2 (0.00), and D3 (1.00). Lastly, A is the average discrimination in the pool with values of A1 = 0.6 and A2 = 1.00.

Overall, Bias and RMSE numbers for each condition in the study are available in Appendix A.

Summary

This section presented the results from the automated test assembly (ATA), the routing mechanism and finally the ca-MST simulations. These results directly target the three research questions described in Chapter I. The four routing and scoring methods were compared against each other in terms of bias and RMSE (see Figure 4.18). The ATA plots, main plots and the tabled values (Appendix A) all described the conditions under various ca-MST design configurations tended to work better with the item pool characteristics. The item pool characteristics questions were answered by the plots describing the effect of the item pool characteristics on the quality of the ca-MST under the four routing and scoring methods.

Overall, the results suggested that the maximum information method was superior to the exposure control method in terms of RMSE. While the IRT method was slightly better than the NC method, the difference was negligible. The ATA tended to work a lot better when the item bank was centered at 0.00 rather than one of the extremes (-1.00 or 1.00); however, measurement recovery was not strongly adversely affected (see Figure 4.7). Even so, there was a visual difference in the RMSE plot when the difficulty was at 0.00 compared to when it was off-center.

Other findings include increases in the module length results in decreases in RMSE. More items taken generally leads to better psychometric accuracy. In addition, when the configuration has more stages, the RMSE dropped due to the increase in items. While these results are not unexpected it’s important information for testing companies desiring a switch to a ca-MST to consider when choosing their configuration.

CHAPTER V

DISCUSSION

This section discusses the implications of the results presented in Chapter IV and places it in the realm of research discussed in Chapters I and II. The results and the implications of the results for each of the three research questions will be discussed. Finally, next steps for further research will be presented.

Impact of Routing and Scoring Methods

This section discusses the implications of the results of the ca-MST simulations and the four routing and scoring methods. This study evaluated the efficacy of four routing and scoring methods in terms of recovery of ability in a ca-MST framework. The four methods were a fully crossed design of two routing/module selection criteria (Maximum Information routing and Exposure Control routing) and two scoring designs (Item Response Theory EAP estimates versus Number-Correct scoring). A full description of these methods and a table of each of the four methods can be found in Chapter I.

Maximum information module selection routes the examinee to the module most informative for the examinee's current estimate of ability. This is probably the most popular module selection criterion for current adaptive testing programs (including CAT). The other module selection criterion could weight the selection criteria by desired population/exposure values. In this study, equal weights were applied to the module

selection criteria. For example in a 1-2-3 ca-MST configuration, 50% of people would be routed to each of the modules in stage 2 and 33.33% would be routed to each of the modules in stage 3.

Maximum information routing tended to do much better than exposure control routing in nearly every condition. The overall results were roughly .02 to .03 better for the maximum information routing as opposed to the exposure control module selection methods. While this study only looked at equal proportions of examinees being sent to each module, other weighting methods could be employed.

Ultimately, maximum information seems to do an adequate job of routing examinees to the appropriate modules. While exposure control did a better job of making sure no module was overexposed. Given that the number of panels/modules in this study were great enough to avoid any potential issue with overexposure, the increased precision would probably be desirable. If on the other hand, the testing program had just a few panels and item exposure was a serious risk, the exposure control module selection method might prevent any particular module from being overexposed. While this would not be statistically optimal, the difference between the exposure control method and maximum information method tended to be less when ca-MST configuration was 1-2-3-4. This finding is probably a result of the amount of items asked overriding potential risk of not sending the examinee to the "most informative" module for the examinees ability estimate.

Item response theory has been the gold standard for scoring and scaling many computer-based assessments of all models. EAP scores tend to be preferred (Thissen &

Orlando, 2001) because of their bias towards the center (as long as the prior is a normal distribution). The efficacy of number-correct scoring within a fully 3PL model versus IRT scoring had not yet been tested. One drawback of using fully IRT 3PL estimates to route and score examinees is that at each step the item parameters would have to be hosted at the test facility.

This study confirmed that IRT scoring provides a trivial amount of extra precision as compared to number correct scoring. In fact, there were a few conditions such as when the difficulty in the item pool was below 0 (D1 = -1.00) in the 1-3 ca-MST configuration that number correct scoring tended to do the same or ever so slightly better than the IRT equivalent. Overall, the difference between IRT and number correct scoring in all iterations and all routing methods was small (roughly .008 for max information routing and .01 for exposure control routing both on the RMSE of theta metric.). These differences are not likely enough to effect the estimation of examinee ability and thus provide nearly equivalent measures of examinee ability.

The implications of these findings suggest that if a testing company wanted to discontinue the online hosting of item parameters and IRT calculations at each step of a ca-MST that this would not adversely affect the estimation of ability for any examinee. In other words, no “harm” would be done to the candidate based on the choice between number correct scoring and traditional IRT practices.

Ultimately, the decision on the routing and scoring should be weighted with the testing vendors needs and goals. For example if exposure is an issue because the number of panels are few, exposure control metrics may be the best decision for that program.

There is a method that can simultaneously balance exposure risk to precision (Luecht & Burgin, 2003). The program should keep in mind; however, that a simple two-stage configuration may not be the best choice for them. If on the other hand, a testing company needs the most precise measurement accuracy and exposure is not a risk because the number of panels is great, this company has more options but should probably choose the maximum information module selection method. Ultimately, the decision between number correct methods and IRT methods are so small that it's unlikely that the ability estimation will be affected. To that end, either the maximum information module selection with IRT-EAP scoring or the maximum information module selection with number correct scoring will result in similar results.

Impact of ca-MST Configuration and Item Pool Variables

This section addresses the results and implications of the second research question. The second research question asked "*Under what conditions do various ca-MST design configurations work best from a psychometric scoring accuracy perspective with respect to the four routing and scoring mechanisms as well as the various item pool characteristics*"? In other words, under which item pool characteristics and ca-MST configuration decisions does the psychometric scoring accuracy tend to be improved by?

One of the more pronounced differences was the RMSE of theta tending to drop as the number of stages increased. Ultimately, the increase in the number of stages meant more items and more items meant increased precision. All of the routing and scoring methods dropped by around 0.10 in RMSE when the number of stages was increased from 2 to 4. Anecdotally from Figure 4.12 the decrease in RMSE seems greater in

switching from the 1-3 to the 1-2-3 than the 1-2-3 to the 1-2-3-4. If the pool is sufficiently large and variable, adding more stages and making the stages have more levels of difficulty seem to help precision.

The other configuration option is also making a large impact on psychometric quality. The RMSE drops when the module length is 20 items instead of 10 items. Regardless of the configuration having 20 items per module improved the fit. The cause for this is exactly the same as increasing the number of stages within the ca-MST configuration, more items means better precision. The effect of increasing the module length was about the same as increasing the number of stages; RMSE's dropped about .10.

The two configuration variables (number of stages & number of items in a module) seemed to confirm the same thing, more is better. More items leads to increased precision that leads to better recovery of ability. This seems to be true regardless of which scoring method is employed. As a result, testing companies should ensure that the proper numbers of items are on the forms to ensure proper precision. While ca-MST tends to gain some efficiency over paper-pencil or fixed length testing, this efficiency does not mean that the test can be 20 questions long. The takeaway from these two findings is that if the item count is low, measurement precision will be low.

Another factor that ended up having perhaps the biggest effect on RMSE/measurement precision was the average discrimination of items in the item pool. This study examined the effect of two different average discrimination parameters, A1=0.6 and A2=1.0. Overall, the means when the average discrimination parameter in

the pool was 1.0 resulted in a much lower RMSE than when the average discrimination parameter in the pool was 0.6. This result again is not terribly unexpected but the magnitude of the drop is very notable. When items have large discrimination parameters they tend to provide more information not only at the location of their maximum information but also more information across the entire score scale. So items that have high discrimination parameters become more valuable at determining the true ability of the examinee. As a result, precision increases as information increases.

As discussed in Chapter I, CAT algorithms struggle with the overuse of high discrimination items and as a result these items tend to get overexposed. Certain techniques such as Weiss (1973) match to *b* or Chang et al. (2001) which added a-blocking to Weiss's match to *b* have had some success in reducing the overall exposure of these items to the testing program. It's important for programs to develop as many of these high discriminating items as possible. While obtaining an average discrimination of 1.0 might be extremely difficult, writing several highly discriminating items particularly with new assessment practices such as Evidence Centered Design (Mislevy et al., 2003) or Assessment Engineering (Luecht, 2006) can assist item writers in developing items that are of high quality and does so in a way that will produce many high quality items. It's important however even in a ca-MST context to avoid overusing discriminating items because item exposure can turn an item into known content that can reduce the discrimination and have other consequences on the item parameters.

The last condition was that of item difficulty. Item difficulty tended to be the least affected for each of the four scoring and routing methods. The routing and scoring

methods tended to perform more or less the same regardless of where the difficulty of the item pool was centered. This study examined three different locations of difficulty (D1 = -1.00, D2 = 0.00, and D3 = 1.00). Overall, the mean RMSE for the difficulty that was centered at 0.00 was lower than the mean RMSEs for RMSEs centered away from 0 (-1.00 and 1.00). Part of this phenomenon could be a result of the way that the examinee ability parameters were generated as they were also centered at 0.00; however, this difference was usually pretty small overall.

Generally speaking, testing companies tend to have their item pool difficulty centered at the most appropriate location for the purpose of their test. A certification/licensure test may need to set their cut score lower so that the majority of test-takers pass the exam. As a result, if their candidate population is normally distributed with a mean of 0.00 most people on average will pass the exam. If the test is an educational test with the need for many cuts/information at multiple points in the scale, the test centered at 0.00 might be more appropriate and may more accurately measure the same population that is centered at 0.00 than an item pool where the mean difficulty of the item pool is off-center. If the purpose of the test is to award additional credentialing/scholarships to a group of highly advanced examinees, an item pool that is right of the center might be more appropriate. While measurement error might be lowest for examinees that are high ability, examinees that are of average to lower ability might contain more measurement error and thus their RMSE's would be a little higher.

A benefit of ca-MST is that while these phenomena should result in much greater standard errors for examinees away from the mean of the item pools when the item pools

are off-center, the effect of this is smaller than any of the other effects almost to the point of not being noticeable. When the number of stages is smaller (e.g., 1-3) the differences in RMSE between the three difficulty means are very small. As the number of stages gets increased the effect of the item pool separation tends to also increase (in this study the 1-2-3-4 was roughly .02 greater RMSE for the off-center mean difficulties).

Control of the item pool and the configuration are under the complete control of the testing company offering the test. The results here should not shock anyone and the results tend to be consistent for all routing and scoring methods. It should not be a surprise that increasing the number of items will reduce residual error in measuring examinee ability, nor that writing good items is a critical component of sound measurement practice. Testing companies need to keep in mind the purpose for the assessment, the overall desired item exposure for their items particularly their items with better discriminations, and the methods that they use to write/produce items and fill their item banks. All of these characteristics of the assessment design process interact with the type of configuration decisions that a testing company can choose in the development of their assessment.

Impact of Item Pool Characteristics on Psychometric Quality of Forms

This section discusses the results and implications of research question number three. The third research question asked, "*how do various item pool characteristics impact the quality of ca-MST forms from a psychometric perspective*"? This research question delves into the relationship between available supplies in the item pool versus

the demands of the ca-MST configurations place on the demands and that impact on panel quality.

In this study the item pools were "repurposed." As a result, the item pools had more items than were necessary. Module item information functions were created using the entire pool. As a result, the ATA engine did the best job it could do given the items that were available to it. Sometimes the resulting MIF's were not exact particularly at the tails of the distribution. This gap between the MIFs and targets were exacerbated when the difficulty was off-center (Dif = -1.00 or Dif = 1.00).

In almost every case the target MIFs were higher than the actual resulting MIFs. This finding could indicate that the target MIFs were targeted a bit too aggressively for the item pool to meet the demands on information found in the item bank.

Tables 4.1 to 4.3 describe the overall fit of the targets to the actual resulting MIFs generated by ATA Panel Builder (Luecht, 2013). The fit of the targets to the actual is dependent on a few factors. First, the k = 20 modules fit a great deal worse than the k = 10 modules. Secondly, the misfit is worse for the 1-2-3 and 1-2-3-4 than the 1-3 design.

One rationale for the misfit is the number of items used versus the available pool. The item targets were developed by breaking the items into clumps equivalent to the number of stages in the configuration. From there the clumps were divided evenly into the modules (for more on this process see Chapter III). Ultimately, each module target function would be composed of the average item in the item pool. If the ATA then uses more than 50% of the items, there might be some "below average" items in the ATA

which could cause some degree of misfit. Table 5.1 describes the % of the pool used by configuration and module length. In each of the configurations, the k = 20 modules use over 60% of the items. This could potentially explain the less than accurate fit that is seen in the k = 20 module MIFs.

Table 5.1

Number of Items Available and Used by ca-MST Configuration and Module Length

Configuration	k=10 it	k=10 it poolsize	% used	k=20 it	k=20 it poolsize	% used
1-3	240	528	45.45%	480	792	60.61%
1-2-3	360	768	46.88%	720	1152	62.50%
1-2-3-4	480	1056	45.45%	960	1584	60.61%

Another issue is that the most of the misfit is in the tails of the target MIFs. The peaks of the target MIFs seem to be nearly or matched very well, but there is a great deal of misfit in the tails. This is particularly true when the difficulty is off-center. There is great amount of misfit in the module tail that is most distant from the center of the item pool.

Dallas et al. (2012) showed the effect on measurement precision when certain exposure criteria were met. As exposure controls were implemented, the effect was that the item pool became stretched to the point where measurement precision was degraded. The same thing applies to creating MIFs if the available information in the item pool isn't enough to cover the demands of the item pool, measurement precision could become

degraded as a result. It's important for testing companies to have item banks capable of supporting their decisions regarding the configuration decisions in the ca-MST.

Comparison of the Baseline versus the Routing/Scoring Methods

This study included a baseline method in the calculations of bias and RMSE. The baseline method was an unconstrained ca-MST that would mimic an item-level CAT and selects the model at each stage that will contribute the maximum measurement precision (e.g., information) at the provisional estimate of examinee ability. In other words it is very related to the maximum information routing; however some constraints are removed. Operationally for a company to deploy the baseline the test driver would need to run IRT statistics in real time, be capable of computing IRT estimates of ability (EAP or MLE) in real time and be able to compute module-level information functions in real time. All of this places burden on the servers and precludes large volumes of examinees testing at once, unless large amounts of computing power are readily available. The two maximum information methods are similar to the baseline and thus their results are extremely similar to the baseline particularly in the 1-3 configuration where they are .001 off on the RMSE metric. The benefit to using either the maximum information IRT routing/scoring method or the maximum information number correct scoring method is that while they essentially perform similar tasks because the ca-MST modules are pre-constructed the scoring rules/decisions are constructed *apriori* and thus only a score lookup table is needed for full routing and decision making. The only IRT scoring that is required is at the end of the examination, which reduces the overall burden on the testing vendors servers.

The lack of constraints on the model may also encourage groups of examinees to game the exam to see as much content as they can. Students in an unconstrained CAT-like ca-MST could jump from the low module on stage 2 to the high module on stage 3. As testing security is a major threat to validity, safeguards against this practice could be employed to stop this behavior.

Further Research

This section describes further avenues of research to explore regarding ca-MST routing and scoring methods. This section touches on additional research topics such as altering the spread of the difficulty parameters, adding more exotic configurations, and experimenting with different routing designs.

One limitation of the study was the underestimation of the bias in a few conditions. Further investigation of those conditions showed a possibility of a lack of convergence (e.g., fixed c's) using default BILOG commands. The RMSE; however, was unaffected by the convergence and properly reflected the accuracy of the condition.

One condition that could be interesting particularly in examining research question #3 about the psychometric quality of the panels based on the available item pool characteristic is to alter the standard deviation of the difficulty parameters. This alteration would change the shape of the distribution of difficulty parameters. If the difficulty parameters are tight, one might assume that the spread of the parameters would change and that could affect the psychometric quality of the ca-MST due to the availability of information being spread over a small area.

Another interesting condition might be adding some more exotic conditions to the study. This study sampled three ca-MST configurations out of a very deep population of potential configurations. A study could sample configurations such as 1-6-6-6-6 or 1-4-5 or others with much larger spreads in difficulty or more stages.

Finally as was mentioned earlier, the exposure control method essentially sent equal numbers of people to the next stage's modules. This would work well for controlling the number of people who saw any given module; however, this design could also be controlled perhaps to route people in a licensure exam to create bins of examinees. For instance, one might have a module for people who are "clear passers," "borderline candidates" and "clear failers." The modules could contain differing levels of multidimensionality or other features to make diagnostic feedback more reliable. If the pass rate is known, different weights could be applied to the ca-MST routing logic to route examinees to particular modules. These weights could be tested and examined for how well they recover ability in a ca-MST design and also how good the resulting diagnostic information is depending on when they entered the "clear fail" stage within the panel.

These are just a few of the many studies that could be done to examine the field of routing and scoring within a ca-MST. These studies will help inform testing company and further the research base of routing and scoring within a ca-MST.

Summary

This section closes the study and describes its impact and how it fits in with the global realm of ca-MST research. To date, the research focus of ca-MST had focused on

comparison to CAT and LFT in terms of efficiency and accuracy (e.g., Jodoin et al., 2006; Luecht, 2000; Luecht et al., 1996; Luecht & Nungester, 1998; Reese et al., 1999; Schnipke & Reese, 1999; Xing, 2001; Xing & Hambleton, 2004). This study confirmed some of the findings of previous studies, most notably the effect that the length of the test has on precision (Jodoin et al., 2006; Luecht et al., 1996; Luecht & Nungester, 1998; Wang et al., 2012). This study took these studies a step further and expanded the literature base of the routing and scoring conditions.

Lord (1980) outlined those six conditions that should be considered before starting a ca-MST. This study covered all of them in some form. The total number of items in the test should be large enough to ensure accurate measurement precision. This study included equal numbers of items in the initial module but found that expanding the number from 10 to 20 items per module tended to result in much better recovery of examinee ability. The difficulty of the initial module was dependent on the overall difficulty in the item bank, easy, moderate, and hard. This study found that the moderate module tended to result in slightly better accuracy but that the precision was largely dependent on other factors including average discrimination in the item pool, number of items in each module, and total items within the panel.

Further, the study examined looking at different number and difficulty of alternative models in each stage, again dependent on the overall difficulty of the item pool. Finally, this study tackled the last two of Lord (1980) seven attributes, which were cut points for routing examinees to modules and methods for scoring stages and each nth stage test.

This study examined two separate methods for setting cut scores: maximum information routing and exposure control routing. Further research could examine altering the weights to determine whether unequal exposure routing might improve examinee ability estimates.

While there had been a few papers (Luecht, Brumfield, et al., 2006) that had brought up the concept of using a number correct scoring algorithm; however, this is the first empirical examination of the effect on measurement accuracy of this method. This study found that number-correct scoring may give up a small amount of precision but other factors are far more influential in determining the overall accuracy in measuring examinee ability.

REFERENCES

Ariel, A., van der Linden, W. J., & Veldkamp, B. P. (2006). A strategy for optimizing item-pool management. *Journal of Educational Measurement, 43*(2), 85–96.

Bejar, I. I. (1991). A methodology for scoring open-ended architectural design problems. *Journal of Applied Psychology, 76*(4), 522.

Bejar, I. I., Lawless, R. R., Morley, M. E., Wagner, M. E., Bennett, R. E., & Revuelta, J. (2003). A feasibility study of on-the-fly item generation in adaptive testing. *The Journal of Technology, Learning, and Assessment, 2*(3). Retrieved from http://ejournals.bc.edu/ojs/index.php/jtla/article/viewFile/1663/1505

Bock, R. D. (1985). *Multivariate statistical methods in behavioral research*. Scientific Software International.

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

Chang, H. H., Qian, J., & Ying, Z. (2001). a-Stratified multistage computerized adaptive testing with b blocking. *Applied Psychological Measurement, 25*(4), 333-341.

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

Chen, S.-K., Hou, L., & Dodd, B. G. (1998). A comparison of maximum likelihood estimation and expected a posteriori estimation in CAT using the Partial Credit

Model. *Educational and Psychological Measurement, 58*(4), 569–595. doi: 10.1177/0013164498058004002

Chen, S-K., Hou, L., Fitzpatrick, S. J., & Dodd, B. G. (1997). The effect of population distribution and method of theta estimation on Computerized Adaptive Testing (CAT) using the Rating Scale Model. *Educational and Psychological Measurement, 57*(3), 422–439. doi: 10.1177/0013164497057003004

Crotts, K., Sireci, S. G., & Zenisky, A. (2012). Evaluating the content validity of multistage-adaptive tests. *Journal of Applied Testing Technology*, *13*(1).

Dallas, A. D., Wang, X., Furter, R., & Luecht, R. M. (2012). *Item pool size, targeted item writing, and panel replication strategies for a 1-3-3 multistage test design.* Presented at the National Council on Measurement in Education, Vancouver, BC, Canada.

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

DeMars, C. (2010). *Item response theory*. New York, NY: Oxford University Press.

Dodd , B. G., & Fitzpatrick, S. J. (2002). Alternatives for Scoring CBTs. In C. N. Mills, M. T. Potenza, J. J. Fremer, & W. C. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 215–234). Mahwah, NJ: Lawrence Erlbaum.

Embretson, S. (1999). Generating items during testing: Psychometric issues and models. *Psychometrika*, *64*(4), 407–433. doi:10.1007/BF02294564

Folk, V. G., & Smith, R. L. (2002). Models for delivery of CBTs. In C. N. Mills, M. T. Potenza, J. J. Fremer, & W. C. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 41–66). Mahwah, NJ: Lawrence Erlbaum.

Gierl, M., & Lai, H. (2011). *The role of item models in automatic item generation.* Paper Presented at the National Council on Measurement in Education Conference, New Orleans, LA.

Gorin, J. S., Dodd, B. G., Fitzpatrick, S. J., & Shieh, Y. Y. (2005). Computerized adaptive testing with the Partial Credit Model: Estimation procedures, population distributions, and item pool characteristics. *Applied Psychological Measurement, 29*(6), 433–456. doi:10.1177/0146621605280072

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston, MA: Kluwer.

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

Holland, P. W. (1990). On the sampling theory roundations of item response theory models. *Psychometrika, 55*(4), 577–601.

Huff, K., Steinberg, L., & Matts, T. (2010). The promises and challenges of implementing evidence-centered design in large-scale assessment. *Applied Measurement in Education, 23*(4), 310–324. doi:10.1080/08957347.2010.510956

Jodoin, M. G., Zenisky, A., & Hambleton, R. K. (2006). Comparison of the psychometric properties of several computer-based test designs for credentialing exams with

multiple purposes. *Applied Measurement in Education, 19*(3), 203–220. doi: 10.1207/s15324818ame1903_3

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

Kolen, M. J., & Tong, Y. (2010). Psychometric properties of IRT proficiency estimates. *Educational Measurement: Issues and Practice*, *29*(3), 8–14. doi:10.1111/j.1745-3992.2010.00179.x

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

Lord, F. N., & Novick, M. P. MR (1968). *Statistical theories of mental test scores.* Reading MA, Addison-Welsley Publishing Company.

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

Luecht, R. M. (2000). *Implementing the Computer-Adaptive Sequential Testing (CAST) framework to mass produce high quality computer-adaptive and mastery tests.* Retrieved from http://www.eric.ed.gov/ERICWebPortal/contentdelivery/servlet/ERICServlet?accno=ED442823

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

Luecht, R. M. (2006). *Assessment engineering: An emerging discipline.* Centre for Research in Applied Measurement and Evaluation, University of Alberta, Edmonton, AB, Canada.

Luecht, R. M. (2009). Adaptive computer-based tasks under an assessment engineering paradigm. In *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing.* Reston, VA: Graduate Management Admission Council.

Luecht, R. M. (2011) *Assessment design and development version 2.0: From art to engineering.* Invited, closing keynote address at the 2011 Annual Meeting of the Association of Test Publishers, Phoenix, AZ.

Luecht, R. M. *Function of Residuals.* Document.

Luecht, R. M. (2012). *Computerized adaptive multi-stage considerations.* Paper Presented at the National Council of Measurement in Education in Vancouver, BC.

Luecht, R. M. (2013a). ca-MST Simulator (software).

Luecht, R. M. (2013v). ATA Panel Builder (software).

Luecht, R. M., Brumfield, T., & Breithaupt, K. (2006). A testlet assembly design for adaptive multistage tests. *Applied Measurement in Education, 19*(3), 189–202. doi:10.1207/s15324818ame1903_2

Luecht, R. M., & Burgin, W. (2003). *Test information targeting strategies for adaptive multistage testing designs.* Annual meeting of the National Council on Measurement in Education, Chicago, IL. Retrieved from http://www.psych.umn.edu/psylabs/catcentral/pdf%20files/lu03-03.pdf

Luecht, R. M., Dallas A. D., & Steed, T. (2010). *Engineering task models: A new way to develop test specifications.* Presented at the National Council on Measurement in Education, Denver, Colorado.

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

Luecht, R. M., & Nungester, R. J. (1998). Some practical examples of computer-adaptive sequential testing. *Journal of Educational Measurement, 35*(3), 229–249. doi:10.1111/j.1745-3984.1998.tb00537.x

Mislevy, R. J., Almond, R. H., & Lukas, J. F. (2003). *A brief introduction to evidence-centered design*. Princeton, NJ: Educational Testing Service.

Paap, M. C. S., & Veldkamp, B. P. (2012). *Minimizing the testlet effect: Identifying critical testlet features by means of tree-based regression.* In: Psychometrics in practice at RCEC. RCEC, Enschede, 74–83. ISBN 9789036533744

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

Pine, S. M. (1977) *Reduction of test bias by adaptive testing.* Paper accessed on http://iacat.org/sites/default/files/biblio/pi77128.pdf

Rasch, G. (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests* (Vol. XIII). Oxford, England: Nielsen & Lydiche.

Reese, L. M., Schnipke, D. L., & Luebke, S. W. (1999). *Incorporating content constraints into a multi-stage adaptive testlet design.* Law School Admission Council Computerized Testing Report. LSAC Research Report Series. Retrieved from http://www.eric.ed.gov/ERICWebPortal/contentdelivery/servlet/ERICServlet?accno=ED467816

Revuelta, J., & Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement, 35*(4), 311–327. doi:10.1111/j.1745-3984.1998.tb00541.x

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. Retrieved from http://www.eric.ed.gov/ERICWebPortal/contentdelivery/servlet/ERICServlet?accno=ED467812

Stocking, M. L., & Lewis, C. (2002). Methods of controlling the exposure of items in CAT. In W. J. van der Linden & G. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 163–182). Springer Netherlands. Retrieved from http://www.springerlink.com/content/g77771n7hx145343/abstract/

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

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

Thissen, D., & Wainer, H. (Eds.). (2001). *Test scoring.* Routledge.

van der Linden, W. J., & Glas, C. A. W. (2010). *Elements of adaptive testing*. New York, NY: Springer.

van Der Linden, W. J. (2005). A Comparison of Item-Selection Methods for Adaptive Tests with Content Constraints. *Journal of Educational Measurement*, *42*(3), 283-302.

van der Linden, W. J., & Veldkamp, B. P. (2004). Constraining item exposure in computerized adaptive testing with shadow tests. *Journal of Educational and Behavioral Statistics, 29*(3), 273–291. doi: 10.3102/10769986029003273

Veerkamp, W. J. J., & Berger, M. P. F. (1997). Some new item selection criteria for adaptive testing. *Journal of Educational and Behavioral Statistics*, *22*(2), 203–226.

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

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24*(3), 185–201.

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

Wald, A. (1947). *Sequential analysis*. 1947 John Wiley & Sons.

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

Way, W. D. (1998). Protecting the integrity of computerized testing item pools. *Educational Measurement: Issues and Practice*, *17*(4), 17–27.

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

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21*(4), 361–375. doi:10.1111/j.1745-3984.1984.tb01040.x

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

Williams, V. S., Pommerich, M., & Thissen, D. (1998). A comparison of developmental scales based on Thurstone methods and item response theory. *Journal of Educational Measurement, 35*(2), 93–107.

Xing, D. (2001). *Impact of several computer-based testing variables on the psychometric properties of credentialing examinations*. Retrieved from http://scholarworks.umass.edu/dissertations/AAI3012196

Xing, D., & Hambleton, R. K. (2004). Impact of test design, item quality, and item bank size on the psychometric properties of computer-based credentialing examinations. *Educational and Psychological Measurement, 64*(1), 5–21. doi: 10.1177/0013164403258393

Yen, W. M. (1984). Obtaining Maximum Likelihood trait estimates from number-correct scores for the Three-Parameter Logistic Model. *Journal of Educational Measurement*, *21*(2), 93-111.

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, NY: Springer. Retrieved from http://www.springerlink.com/content/r173k1747l175v63/abstract/

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). *BILOG-MG 3 for Windows: Multiple-group IRT analysis and test maintenance for binary items* [Computer software]. Skokie, IL: Scientific Software International, Inc.

APPENDIX A

BIAS AND RMSE BY CONFIGURATION TABLES

Table A.1

Bias by Configuration

Config	BL_Bias	SD(B_LBias)	MI_IRT	SD(MI_IRT)	MI_NC	SD(MI_NC)	EC_IRT	SD(EC_IRT)	EC_NC	SD(EC_NC)
Overall	-0.006	0.015	-0.007	0.016	-0.007	0.015	-0.007	0.017	-0.006	0.015
C1	0.001	0.009	0.000	0.010	-0.001	0.010	0.001	0.016	0.000	0.014
C2	-0.011	0.016	-0.011	0.016	-0.011	0.015	-0.011	0.016	-0.010	0.015
C3	-0.010	0.017	-0.011	0.017	-0.010	0.016	-0.011	0.017	-0.010	0.015
C1P1D1A1	0.002	0.010	0.001	0.011	0.002	0.011	0.004	0.009	0.004	0.010
C1P1D1A2	0.003	0.010	0.002	0.010	0.003	0.011	0.006	0.011	0.006	0.011
C1P1D2A1	0.000	0.009	-0.001	0.008	-0.002	0.008	-0.002	0.009	-0.002	0.008
C1P1D2A2	-0.001	0.006	0.000	0.006	-0.001	0.006	-0.001	0.006	-0.002	0.006
C1P1D3A1	0.009	0.008	0.007	0.008	0.005	0.008	0.001	0.008	0.001	0.008
C1P1D3A2	0.001	0.009	0.001	0.009	0.000	0.009	-0.003	0.010	-0.004	0.010
C1P2D1A1	0.001	0.011	-0.007	0.017	-0.006	0.017	0.015	0.043	0.013	0.036
C1P2D1A2	0.004	0.007	0.004	0.007	0.004	0.007	0.007	0.008	0.006	0.008
C1P2D2A1	-0.004	0.011	-0.006	0.012	-0.006	0.012	-0.006	0.012	-0.006	0.012
C1P2D2A2	0.000	0.005	0.000	0.006	0.000	0.006	-0.001	0.006	-0.001	0.006
C1P2D3A1	0.003	0.008	0.001	0.008	0.000	0.009	-0.002	0.009	-0.003	0.009
C1P2D3A2	-0.005	0.006	-0.005	0.006	-0.006	0.007	-0.010	0.006	-0.009	0.006
C2P1D1A1	0.001	0.009	-0.001	0.009	0.001	0.009	0.003	0.010	0.004	0.010
C2P1D1A2	-0.002	0.011	-0.002	0.011	-0.001	0.010	0.001	0.010	0.001	0.010
C2P1D2A1	-0.001	0.008	-0.001	0.008	-0.001	0.009	-0.002	0.009	-0.001	0.009

Table A.1

(Cont.)

Config	BL_Bias	SD(B_LBias)	MI_IRT	SD(MI_IRT)	MI_NC	SD(MI_NC)	EC_IRT	SD(EC_IRT)	EC_NC	SD(EC_NC)
C2P1D2A2	-0.004	0.011	-0.004	0.011	-0.004	0.010	-0.005	0.011	-0.004	0.010
C2P1D3A1	0.003	0.007	0.002	0.007	0.000	0.006	-0.005	0.005	-0.004	0.005
C2P1D3A2	-0.003	0.011	-0.003	0.010	-0.003	0.011	-0.007	0.011	-0.008	0.011
C2P2D1A1	-0.023	0.012	-0.022	0.012	-0.021	0.012	-0.020	0.012	-0.019	0.012
C2P2D1A2	-0.023	0.009	-0.023	0.009	-0.022	0.009	-0.018	0.008	-0.015	0.008
C2P2D2A1	-0.017	0.013	-0.018	0.012	-0.017	0.012	-0.018	0.012	-0.017	0.012
C2P2D2A2	-0.041	0.012	-0.042	0.012	-0.038	0.012	-0.043	0.013	-0.040	0.013
C2P2D3A1	-0.006	0.012	-0.008	0.011	-0.009	0.011	-0.010	0.010	-0.008	0.010
C2P2D3A2	-0.011	0.012	-0.013	0.012	-0.012	0.012	-0.008	0.011	-0.006	0.011
C3P1D1A1	-0.003	0.010	-0.004	0.010	-0.002	0.010	-0.001	0.010	0.000	0.010
C3P1D1A2	0.005	0.007	0.005	0.007	0.005	0.006	0.007	0.007	0.006	0.007
C3P1D2A1	-0.001	0.009	-0.002	0.009	-0.003	0.008	-0.003	0.009	-0.003	0.009
C3P1D2A2	0.006	0.007	0.005	0.007	0.004	0.007	0.004	0.007	0.003	0.007
C3P1D3A1	-0.002	0.010	-0.003	0.010	-0.004	0.011	-0.010	0.011	-0.008	0.010
C3P1D3A2	0.002	0.009	0.002	0.009	0.002	0.009	-0.005	0.010	-0.005	0.010
C3P2D1A1	-0.020	0.006	-0.020	0.006	-0.019	0.006	-0.015	0.005	-0.014	0.006
C3P2D1A2	-0.034	0.010	-0.034	0.010	-0.032	0.010	-0.026	0.010	-0.022	0.010
C3P2D2A1	-0.014	0.008	-0.018	0.007	-0.017	0.007	-0.017	0.008	-0.016	0.008
C3P2D2A2	-0.044	0.008	-0.045	0.008	-0.040	0.008	-0.047	0.008	-0.042	0.008
C3P2D3A1	-0.002	0.010	-0.006	0.011	-0.006	0.011	-0.010	0.010	-0.008	0.010
C3P2D3A2	-0.010	0.012	-0.012	0.011	-0.012	0.011	-0.009	0.011	-0.006	0.011

Table A.2

RMSE by Configuration

Config	BLRMSE	SD(BLRMSE)	MI_IRT	SD(MI_IRT)	MI_NC	SD(MI_NC)	EC_IRT	SD(EC_IRT)	EC_NC	SD(EC_NC)
Overall	0.344	0.086	0.346	0.087	0.354	0.086	0.357	0.090	0.366	0.089
C1	0.401	0.085	0.402	0.086	0.403	0.087	0.415	0.088	0.416	0.088
C2	0.337	0.073	0.339	0.074	0.346	0.076	0.348	0.076	0.356	0.078
C3	0.294	0.064	0.298	0.067	0.313	0.069	0.308	0.070	0.325	0.075
C1P1D1A1	0.515	0.005	0.516	0.006	0.516	0.005	0.519	0.004	0.518	0.004
C1P1D1A2	0.375	0.003	0.375	0.004	0.375	0.004	0.386	0.003	0.387	0.003
C1P1D2A1	0.510	0.004	0.511	0.005	0.512	0.004	0.513	0.005	0.513	0.004
C1P1D2A2	0.374	0.009	0.376	0.014	0.377	0.013	0.380	0.014	0.380	0.011
C1P1D3A1	0.541	0.002	0.541	0.002	0.546	0.002	0.563	0.003	0.565	0.004
C1P1D3A2	0.423	0.007	0.423	0.007	0.428	0.006	0.476	0.006	0.487	0.006
C1P2D1A1	0.403	0.042	0.406	0.048	0.405	0.049	0.409	0.054	0.409	0.055
C1P2D1A2	0.279	0.001	0.280	0.001	0.280	0.002	0.289	0.002	0.289	0.002
C1P2D2A1	0.384	0.003	0.384	0.003	0.384	0.003	0.385	0.003	0.385	0.003
C1P2D2A2	0.270	0.002	0.270	0.002	0.270	0.002	0.273	0.002	0.274	0.002
C1P2D3A1	0.415	0.005	0.415	0.005	0.416	0.005	0.429	0.005	0.430	0.006
C1P2D3A2	0.322	0.003	0.322	0.003	0.323	0.003	0.354	0.003	0.360	0.004
C2P1D1A1	0.440	0.002	0.444	0.002	0.452	0.003	0.446	0.003	0.457	0.004
C2P1D1A2	0.318	0.002	0.318	0.002	0.327	0.002	0.329	0.002	0.346	0.002
C2P1D2A1	0.438	0.004	0.441	0.004	0.452	0.005	0.441	0.004	0.451	0.004
C2P1D2A2	0.309	0.003	0.310	0.003	0.317	0.003	0.313	0.003	0.323	0.003
C2P1D3A1	0.458	0.003	0.460	0.003	0.469	0.004	0.478	0.002	0.488	0.003
C2P1D3A2	0.362	0.005	0.362	0.006	0.371	0.006	0.404	0.006	0.418	0.006
C2P2D1A1	0.332	0.003	0.333	0.003	0.338	0.002	0.336	0.002	0.341	0.003

Table A.2

(Cont.)

Config	BLRMSE	SD(BLRMSE)	MI_IRT	SD(MI_IRT)	MI_NC	SD(MI_NC)	EC_IRT	SD(EC_IRT)	EC_NC	SD(EC_NC)
C2P2D1A2	0.250	0.002	0.251	0.002	0.256	0.002	0.260	0.002	0.266	0.002
C2P2D2A1	0.316	0.002	0.318	0.002	0.323	0.002	0.316	0.002	0.320	0.002
C2P2D2A2	0.228	0.003	0.229	0.002	0.233	0.003	0.229	0.003	0.233	0.003
C2P2D3A1	0.339	0.003	0.341	0.003	0.344	0.004	0.348	0.003	0.353	0.004
C2P2D3A2	0.261	0.004	0.262	0.004	0.264	0.004	0.274	0.005	0.281	0.005
C3P1D1A1	0.386	0.003	0.393	0.002	0.410	0.004	0.397	0.002	0.417	0.004
C3P1D1A2	0.273	0.001	0.274	0.001	0.295	0.002	0.287	0.002	0.318	0.002
C3P1D2A1	0.380	0.001	0.390	0.002	0.410	0.004	0.388	0.002	0.405	0.002
C3P1D2A2	0.266	0.002	0.269	0.002	0.297	0.004	0.270	0.002	0.298	0.002
C3P1D3A1	0.406	0.003	0.413	0.004	0.433	0.004	0.432	0.003	0.462	0.044
C3P1D3A2	0.315	0.002	0.316	0.002	0.329	0.003	0.363	0.003	0.391	0.005
C3P2D1A1	0.290	0.002	0.294	0.002	0.304	0.002	0.297	0.001	0.306	0.002
C3P2D1A2	0.219	0.004	0.221	0.004	0.231	0.004	0.230	0.004	0.238	0.004
C3P2D2A1	0.273	0.002	0.279	0.003	0.289	0.005	0.275	0.003	0.282	0.003
C3P2D2A2	0.200	0.002	0.202	0.002	0.215	0.003	0.201	0.002	0.213	0.003
C3P2D3A1	0.297	0.003	0.301	0.003	0.309	0.004	0.310	0.003	0.317	0.002
C3P2D3A2	0.227	0.004	0.229	0.004	0.234	0.004	0.243	0.005	0.256	0.006