WANG, XINRUI, Ph.D. An Investigation on Computer-Adaptive Multistage Testing Panels for Multidimensional Assessment. (2013)
Directed by Dr. Richard M Luecht. 89 pp.

The computer-adaptive multistage testing (ca-MST) has been developed as an alternative to computerized adaptive testing (CAT), and been increasingly adopted in large-scale assessments. Current research and practice only focus on ca-MST panels for credentialing purposes. The ca-MST test mode, therefore, is designed to gauge a single scale. The present study is the first step to investigate ca-MST panels for diagnostic purposes, which involve the assessment of multiple attributes in the same test.

This study employed computer simulation to compare multidimensional ca-MST panels and their unidimensional counterparts, and to explore the factors that affect the accuracy and efficiency of multidimensional ca-MST. Nine multidimensional ca-MST panel designs – which differed in configuration and test length – were simulated under varied attribute correlation scenarios. In addition, item pools with different qualities were studied to suggest appropriate item bank design.

The comparison between the multidimensional ca-MST and a sequential of unidimensional ca-MST suggested that when attributes correlated moderate to high, employing a multidimensional ca-MST provided more accurate and efficient scoring decisions than several unidimensional ca-MST with IRT scoring. However, a multidimensional ca-MST did not perform better than its unidimensional counterpart with MIRT scoring. Nevertheless, multidimensional panels are still promising for diagnostic purposes given practical considerations.

The investigation on multidimensional ca-MST design indicated the following: Higher attribute correlation was associated with better scoring decision because more information carried by a correlation matrix was available for estimation. This held true across all item pool conditions. An optimal item pool would be the one that was discriminative, appropriately located and specifically designed for a configuration. The accuracy and efficiency of a multidimensional ca-MST panel would be diminished if its item pool was too easy, or not informative. According to the results, the 1-2-3 configuration design was most promising. In terms of test length, an appropriate decision would largely depend on the attribute correlation and the item pool characteristics.

AN INVESTIGATION ON COMPUTER-ADAPTIVE MULTISTAGE TESTING PANELS FOR MULTIDIMENSIONAL ASSESSMENT

by

Xinrui Wang

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
2013

Approved by

_____
Committee Chair

To Xiao, Sophie, and my family.

APPROVAL PAGE

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

Committee Chair _____

Committee Members _____

_____

_____

_____

_____
Date of Acceptance by Committee

_____
Date of Final Oral Examination

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER I

INTRODUCTION

The fast development in computer capability and software engineering, along with the increasing accessibility of large computer labs in educational sites and testing centers have facilitated the wide-spread employment of CAT in current testing industry. Many large-scale standard tests have switched from paper and pencil (P&P) formats to CAT aiming for different purposes, such as Graduate Management Admission Test (GMAT) for graduate school admission, National Council Licensure Examinations (NCLEX) for registered nurse licensure, and COMPASS series of tests that provides placement and diagnostic testing for English as Second Language (ESL) students.

In a CAT approach, items are customized to examinees during testing process. The ability location of an examinee is estimated by the first few items. The following item or item set is selected from item bank to maximize or minimize a criterion related to the measurement of location. After administering the selected item/item set, examinees' ability location is updated and used to select the new item/item set. These processes continue until certain measurement accuracy is achieved or until reaching the maximum test length. The main advantage of CAT is that it achieves desired measurement accuracy with shorter tests than conventional forms. In addition, CAT does not require test administration for all examinees at the same time because each test taker's form is customized. The employment of computers as the vehicle of delivering tests also brings

in advantages such as immediate scoring feedback, greater standardization of test administration, collection and storage of various types of information, better control of test security and adopting innovative item types (Chalhoub–Deville & Deville, 1999).

In a large-scale assessment, quality control of test forms before test delivery is indispensable in that the corresponding scoring procedure often involves high stake decisions that are affecting test takers as individuals as well as school programs. This task becomes difficult when CAT is employed, due to the fact that test forms would not be assembled until the end of a CAT administration. To overcome this shortcoming of CAT while benefitting from the psychometric efficiency of adaptive testing, computer adaptive multistage testing (ca-MST), or computer-adaptive sequential testing (CAST) was introduced to better control test quality and exposure rate. Analogous to traditional computer-adaptive testing (CAT), ca-MST involves adaptive selection of items according to examinee's ability. However, instead of adapting every item, the unit of ca-MST is a testlet, or a test "module" termed in Luecht & Nungester (1998). That is, a group of items are adapted. Test modules are preassembled in a test panel with several stages (Luecht & Nungester, 1998). At the beginning of a ca-MST test, a test panel is randomly assigned to the test taker, and the adapting happens within the panel. Because all the modules in a panel are known before hand, all possible test forms that a panel can generate are also predetermined. Although ca-MST is also a mode of computer adaptive testing, in this paper, CAT only refers to the item-level computer adaptive testing.

Luecht (2000) suggested some practical advantages of ca-MST over CAT. First, because an adaption happens within a panel, the complete test forms can be reviewed and

well-controlled when panels are assembled. Second, the data management and processing loads are minimized given the simplicity of scoring and routing mechanism. Third, ca-MST provides straightforward ways of dealing with item and test exposure risk. An additional benefit is that ca-MST allows test takers to review and change answers within models during the test administration. Currently, ca-MST has been adopted by Certified Public Accountants (CPA) exam, as well as the Graduate Record Examination (GRE).

Tests for Diagnostic Purposes

Diagnostic assessment provides a profile of strength and weakness for each examinee. In the education field, diagnostic assessment can help teachers identify problems and facilitate further instructions; in the licensure field, test candidates, especially those who failing the exam could benefit from tests that have diagnostic function built in, which sheds some light on further test preparation directions.

The purpose of an assessment should guide the test design process, which involves establishing the linkage among three models: 1) a theoretical construct model; 2) a test development model; 3) a psychometric scoring model (Luecht, 2003). When the main concern of an assessment is to provide diagnostic information regarding test candidates' strength and weakness, the underlying construct of interest would be multivariate representing separate meaningful skill components. In this case, a multidimensional scoring model would be appropriate to serve the scoring purpose and

provides each examinee with a profile with scoring on each of the hypothesized skill constructs.

During past decades, many studies have devoted to psychometric models for obtaining diagnostic information. The most commonly used method in public school system is reporting raw scores for each of the subscales (Stone, Ye, Zhu, & Lane, 2009; Sinharay, Puhan, &Haberman, 2011). To make the subscores more reliable, previous research has established regressed augmented subscores that added value in reporting over raw subscores (Haberman, 2008; Sinharay, 2010; Stone, et al., 2009). Another subscoring method is to produce zero/one score that represents nonmastery or mastery of each skill. A family of models that relate observed responses to mastery status of underlying ability is the diagnostic classification models (DCM). Although DCM has attracted a lot of attention during past several decades, its application is still limited in practice. The most popular psychometric scoring models in the current testing industry are the IRT models. Separate IRT scoring on subscales is the simplest way to report diagnostic information. However, when more than one skill is measured in an item, an extension to IRT model – the multidimensional IRT (Reckase, 1972) could be more appropriate to capture characteristic of the multivariate latent space.

Considerations for Diagnostic ca-MST Design

In a diagnostic test, the goal is to extract as much information on the multiple abilities required to solve the test items as possible (Mulder & van der Linden, 2009). The adaptive testing format is therefore more efficient and reliable in providing such

information. Previous researches have addressed computer adaptive testing procedures in the multidimensional IRT framework (e.g., Luecht, 1996; Mulder & van der Linden, 2009; Segall, 1996). Recent studies also tackled multidimensional CAT in the DCM framework (e.g., Cheng, 2009; Wang, Chang, & Douglas, 2011). However, there is a lack of literature on multidimensional ca-MST procedures except Luecht (2012). Most of the previous ca-MST researches have focused on the certification purpose, in which employing unidimensional IRT model is sufficient. Yet ca-MST designs have not been adopted for diagnostic purposes. To set up a diagnostic ca-MST, considerations need to be addressed regarding both the psychometric models and the ca-MST design itself as listed below. In this study, each item is supposed to measure one latent trait (simple structure).

Employment of Appropriate Scoring Methods

Because of the simple structure assumption, a unidimensional or a multidimensional scoring model may be appropriate. When a unidimensional model is applied, subscores can be obtained by scoring each subscale separately. If a multidimensional framework is employed, more than one latent scale can be calibrated simultaneously. Each model has its advantages and disadvantages depending on the testing context. Therefore, an appropriate scoring method needs to be explored.

Besides the final scoring, ca-MST panels need an extra scoring procedure for routing. Test candidates can be routed to the next stage by scoring the current stage, or by maximizing or minimizing certain statistic criterion, which circumvents the scoring

step. In the unidimensional framework, employing the number of correct (NC) scoring strategy for routing can be very efficient. In the multidimensional world, on the other hand, when scoring each stage, the NC scoring strategy may or may not work, and cut scores may need to be carefully decided for each dimension. If certain statistic criterion is applied, the adaptive response time should be seriously considered especially with high dimensionality tests.

Multidimensional Ca-MST Configuration Design

A typical ca-MST configuration design involves the determinations on the number of adaptive stages and the number of modules within each stage. The number of module within each stage reflects the level of adaptation. Although higher adaptation is desirable, multiple dimensions, along with fine adaptation may result in a design that is too complicated. For example, in the second stage of a unidimensional 1-3 ca-MST design, three modules contain testing items of high, medium and low difficulty that are most informative for high, medium and low ability group respectively. When shifting to a k dimensional space, different ability groups are characterized by more than one skill or trait, and therefore would be defined according to 3^k ability compositions if each ability scale were divided into three levels. To accommodate for each group, 3^k distinct modules are needed, which generally results in too many modules within an adaptive stage.

With the increase in number of dimensions, the desired number of module in each adaptive stage increases geometrically. On the other hand, as the number of adaptive stage increases, the number of possible routes within a panel becomes huge and hence led

to tremendous demand of test form review work as well as item bank size, which could be far from realistic.

It is worth noting that the correlation among measured traits determines the distribution of ability groups. If the skills are reasonably correlated, some ability categories would have sparse population and could be less of considered. In addition, certain routes become infeasible and thus may be eliminated. When designing a multidimensional ca-MST test, factors that take care of dimensionality should be addressed while reasonable constraints need to be developed for efficiency and economic concerns.

Providing Reliable Scores

Within the IRT framework, the reliability of a score depends on the information the test provides at the ability location corresponding to that score. In an adaptive test, the amount of information at each scale point is largely affected by the item pool. If an item pool is not able to provide enough informative items within certain ability regions, reliable score estimations are not expected given limited test length.

However, item pools often, if not always, cannot be developed ideally. A common problem is that items of medium difficulty often take the majority part of an item pool. This generally results in a shortage of informative items for groups at the two ends of an ability scale, and therefore the reliability of their scores is subject to doubt.

Choosing a scoring method could also influence the reliability of subscores when the test is multidimensional, especially when measured traits are correlated. Due to the

inter-correlation among dimensions, the measure on one dimension can provide information for other correlated dimensions. In developing a scheme for multidimensional ca-MST, the scoring model needs to be carefully chosen so that the information carried by inter-trait correlations can be exploited to yield more reliable scores.

Test length is another factor that influences reliability. When multiple traits are measuring the same test, there is no guarantee that each trait is tackled by enough number of items. Therefore, a strategy that ensures the amount of information at each ability level needs to be developed to obtain reliable scores.

Purpose and Rationale of Research

The purpose of this study is to explore the complexity in developing a scheme for diagnostic ca-MST. Specifically, it discusses the effect of the correlation among measured traits and the item bank characteristics on the accuracy and efficiency of multidimensional ca-MST panels.

When multiple traits are involved in the same test, their inter-correlation is essential. Multidimensional ca-MST design may be unnecessary when traits are highly correlated so that a single dimension can capture most of the information. On the other extreme, when traits are uncorrelated, measuring multiple traits simultaneously may not outperform separate assessment on each of the traits. In this study, the effect of correlation matrix will be evaluated through scoring precision and item bank requirement.

When the measured traits are reasonably correlated, the proficiency level on one latent trait can suggest the level of other related traits in some degree. That being said, not only the items measuring a particular trait but also the items that measure related trait(s) can provide information toward ability estimations. Therefore, we may not need highly informative items on each trait to achieve scoring precision. In this case, the demand for high quality item pool can be loosen. This study considers a diversity of item pool characteristics, explores its effect and delves into its interaction with the magnitude of latent trait correlation.

CHAPTER II

REVIEW OF THE LITERATURE

The literature review mainly focuses on the design of computer-adaptive multistage testing. Zenisky, Hambleton and Luecht (2010) suggested seven design considerations for selecting a panel configuration *n* stages:

- the total number of items in the test;
- the number of stages needed;
- the number of items in the initial module versus at each;
- the number of difficulty-level modules available for branching at each stage;
- the distributions of difficulty and discrimination in the initial module versus at later stages;
- cut-points or mechanisms for routing examinees to modules with the panel; and
- method for scoring modules through the n^{th}-stage of the test.

As pointed out by Zenisky et al. (2010) and Wang, Fluegge and Luecht (2012), other considerations include the examinee proficiency or ability population distribution, the extent of test information overlap for modules within stages if modules are constrained by the restricted statistical characteristics of the item bank, whether modules should be fixed on a specific panel or randomly selected from difficulty-based "bins" at

each stage, whether content balancing is carried out at the module or total test levels, the choice of test assembly algorithms, heuristics and software, the size and quality of the item bank, how test information is distributed across stages (in aggregate, at the test level), the placement of cut-scores for pass–fail decisions, the issue of test form and item review by test development and/or subject-matter experts, and likely item-exposure rates/risk and associated issues. Apparently, it is impossible to address all these issues in one study. Related to the current topic, we will focus our discussion on panel configuration design, panel assembly, routing strategy, item bank, and scoring. Also, the technical aspects on ca-MST panels for multidimensional assessment are presented.

Panel Configuration Design

The design of panel configuration includes the decision on the number of stages and the number of difficulty-level modules available for branching at each stage. The exact number of stages is a test design decision affected by the extent of desired content coverage and measurement precision (Zenisky et al., 2010). More stages in a panel simply indicate more chances of adaptation and less risk of measurement error if any accidental responses occur. For example, when an above-average examinee unexpectedly slips in the first stage, he/she will be routed to an easy module, which is not likely to correctly reflect his/her ability level. Adding a third stage offers another chance to route the examinee to a more informative module. Similarly, the number of modules in a stage affects the measurement precision. More modules per stage with a variety of

difficulty levels make a test more adaptable to a wider range of examinee proficiency levels.

In general, having more stages and using testlets of more varied difficulty per stage allow for greater adaptation (Luecht and Bergin, 2003). However, as mentioned in Zenisky et al.'s (2010) paper, higher adaptable stage requires more easy items and hard items to build the MST modules, which is demanding for the item pool construction. Another practical issue needs to consider is the panel review load. More stages and more adaptable stages generally result in the increase of possible testing routes, which means more pre-determined test forms for review.

In the pioneer work of ca-MST (Luecht & Nungester, 1998), 1-3, 1-3-3, and 1-3-5 design were studied in a medical exam context. The 1-2-2 design was discussed in the language testing context (Luecht, 2003). This design has been employed in the Uniform Certified Public Accountant exam. Also, the 1-3-3 design is one of the several that is mostly studied (e.g. Hambleton & Xing, 2006; Luecht & Nungester, 1998; Wang et al., 2012) and considered as promising in practice. Lord (1980), too, suggested that two or three stages and three or four modules at each stage would likely suffice in practice.

Panel Assembly

The assembly of ca-MST occurs at the panel level (Luecht & Nungester, 1998), which includes the assignment of items to modules and modules to stages within panels. The ca-MST test assembly focuses on assembling panels and modules to consistently match multiple statistical targets and content constraints (Luecht & Nungester, 1998).

There are two strategies of panel assembly according to Luecht & Nungester (1998): the Top-Down strategy that specifies statistical targets and content constraints for each of several primary routes or pathways through the ca-MST panel; the Bottom-Up strategy in which assembly targets are specified for each of the modules in the panel. Luecht and Nungester (1998) illustrated the process of panel assembly using both strategies. Compared to the Bottom-Up strategy, which straightforwardly sets qualitative and quantitative constrains for each module, the Top-Down assembly requires more constrain specifications and needs more efforts.

One most used indication of statistical targets is the IRT target test information function (TIF). A target TIF indicates the amount of test information desired across the latent proficiency scale, which can be specified for modules or for particular combinations of modules. Birnbaum (1968) demonstrated the reciprocal relationship between the estimation error variance of ability θ and the test information function as

$$v(\hat{\theta}|\theta) = \frac{1}{\sum_{i=1}^{n} \frac{[P(\theta;\xi_i)']^2}{P(\theta;\xi_i)[1-P(\theta;\xi_i)]}} = \frac{1}{\sum_{i=1}^{n} I(\theta;\xi_i)}, \quad (1)$$

where $v(\hat{\theta}|\theta)$ is the conditional error variance of $\hat{\theta}$ at ability location θ, $P(\theta;\xi_i)$ is an IRT function that models the probability of a correct response given ability θ and item parameters ξ_i, and $I(\theta;\xi_i)$ is the information of an item at location θ which is addable across the test. Therefore, by determining the amount of estimation error we are willing to tolerate, the target TIF can be determined. Lord (1980), Luecht (1992), and van der Linden and Luecht (1995) all provide some techniques for generating target TIF.

Although they aim at deriving target TIFs for fix-length tests, these strategies might be borrowed in the ca-MST context with some modifications.

Luecht & Burgin (2003) suggested three goals in target TIF, including 1) to help guarantee that the IRT test information functions provide measurement precision where it is most needed for critical decisions and score- reporting purposes; 2) to derive targets that make it feasible to actually produce large numbers of content- balanced MST testlets; and 3) to achieve a desired level of conditional exposure of test materials in the examinee population for each constructed panel. To achieve these goals, they proposed the conditional information targeting (CIT) strategy to find appropriate module TIFs and illustrated the strategy in a 1-2 panel design example given an item bank. Three points where three modules maximize measurement precision were first identified on the proficiency scale. N (N was the number of desired panels to be built) non-overlapping testlets that maximize information at each point were then built and the TIF were averaged to obtain the provisional TIF. This process was first described as the approximate maximum information (AMI) in Luecht (2000). After provisional TIFs were sketched out, a two-step process was taken: 1) find the intersection of two TIFs of the second stage; 2) move the maximum precision points of stage-two modules until their intersection point align with the maximum precision point of the stage one module. It is worth noting that the CIT strategy provides an approach to find realistic target TIFs in practice since item bank almost always affects the feasibility of ca-MST panel assembly, which will be discuss in more details later in this section.

In general, the key to generating targets is to focus on the primary pathways within the panel (Luecht, 2000). As suggested by Luecht and Nungester (1998), as more branching modules appear in a later stage, the standard deviation of item difficulty in a single module should decrease, indicating improved adaptivity. Reflecting item difficulty on target TIFs, the shape of target TIFs becomes sharper in later stages.

Once the target TIFs are settled and the content constrains are specified, automated test assembly (ATA) can be carried out to assign items to each module and construct panels. ATA problems can be solved by linear programming algorithms, network-low approximation, or constructive heuristics. The in-depth discussion on these algorithms is beyond the scope of this study. In ca-MST studies, the normalized weighted absolute deviation (NWAD) heuristic proposed by Luecht (1998) is often used (e.g. Hambleton & Xing, 2006; Luecht & Nungester, 1998; Zheng, Nozawa, Gao, & Chang, 2012), which deals with both statistic targets and content constraints of test specifications.

Routing Strategy

Routing strategy refers to the approach of selecting future module given examinees' performance on previous one(s). Routing an examinee from one stage to the next is analogous to the item selection process in a CAT, which chooses an item that is optimal given examinees' provisional proficiency estimation. In ca-MST, modules are scored cumulatively to obtain provisional score estimates for selecting the next module. Scoring options include using number-correct (NC) scoring, cumulative weighted NC,

and IRT-based provisional proficiency scores such as maximum likelihood estimates (MLE) or estimated a prior (EAP) estimates (Zenisky et al., 2010). Although IRT scoring is certainly reasonable, Luecht and Nungester (1998) demonstrated that NC scoring is probably sufficiently accurate for the purpose of module selection, while keeping the scoring procedure and data processing simple.

Once a scoring method is chosen, the next step is to determine cut scores for making routing decisions. Cut scores present as upper bounds and lower bounds for each module. For example, in a 1-3 ca-MST design, routing an examinee from Stage 1 to Stage 2 requires two cut points X_L and X_H. If the provisional score of the examinee in Stage 1 is lower than X_L, he/she will be route to an easy module; if his/her score is higher than X_H, he/she will be route to an difficult module; if the examinee's score falls between X_L and X_H, he/she will be route to an medium difficult module.

Two approaches for determining routing points were described in Luecht, Brumfiled & Breithaupt (2006): the approximate maximum information (AMI) method, and the defined population intervals (DPI) method. Under the former method, AMI approach (Luecht, 2000) was first used to find empirical target TIFs. The TIF of a previous administered module was added to the TIFs of current alternative modules respectively, and these adjacent cumulative TIFs were compared to each other. The intersection of adjacent cumulative TIFs can then be found as the routing point. This method is analogous to the maximum information criteria used in CAT, which selects the module that provides maximum information given the provisional location of an

examinee. Notice that because two panels are very likely to have different TIFs, the cut points for different panels should correspondingly differ.

The second method can be used to implement a policy that specifies the relative proportions of examinees in the population expected to follow each of the primary routes through the panel. In the previous 1-3 ca-MST design circumstance, if we would like the test population to be equally divided to take Easy, Medium and Difficult module in the second stage, the scores of 33% and 67% percentiles would be the cut points. Assuming the population is normally distributed, the routing points would be -0.44 and 0.44.

Although the above two methods refers to the IRT scale, these determined IRT routing points can be transferred to number of correct scores. Given a particular cut point θ_c, and the IRT item parameters for a set of k modules administered up to that point, ξ_i, i =1,…, n_j, j =1, …, k, the corresponding estimated true-score point is

$$X_c = \sum_{j=1}^{k} \sum_{i=1}^{n_j} P(\theta_c; \xi_i). \quad (2)$$

The true score can be further rounded to approximate a number-correct score that can be used for routing decisions (Luecht et al., 2006).

Item Bank

In the test assembly examples of Luecht and Nungester (1998), item bank has been suggested to affect the successfulness of ATA process. If item supply were not sufficient within desired proficiency range, the specifications of ATA would not be met. Xing and Hambleton (2004) also commented that the best test design available cannot

compensate for item lacking in content validity and desirable statistical properties to construct the examination of interest. Jodoin, Zenisky and Hambleton's study (2006) designed a three-stage ca-MST given an operational item bank used for fixed form tests. The results showed that the ca-MST design produced results that were comparable to the previous fixed forms but certainly not better. In a recent large-scale comparative study of ca-MST (Wang, Fluegge, & Luecht, 2012), 25 different panel designed were compared under two item pool conditions: 1) using the existed pool for fixed form tests; 2) design an "optimal" pool for ca-MST. The results, not surprisingly, confirmed that the quality of item bank was perhaps the primary factor that impacted the quality of almost any ca-MST panel design. Using an item pool designed specifically for ca-MST dramatically improved scoring accuracy.

Ultimately, the item bank (the supply) needs to reflect the demands for measurement precision along the ability continuum. For example, if a credentialing examination needs to distinguish primarily among high performing examinees, the corresponding item bank must contain a large number of very difficult items (Wang et al., 2012). Wang et al. (2012) also noted that each ca-MST panel configuration required a potentially different item bank that was optimal for that design. There is not a single item bank that is optimal for all ca-MST designs. In general, item bank provides information supply for test assembly. Statistically, an "ideal" bank for ca-MST would be the one that is sufficiently large in size, and contains items that well target desired difficulty region and discriminate properly.

It is clear that larger item bank provides more freedom in test assembly, yet unreasonable large item pool size simply adds on the burden of item writing cost. In any case, the demand in item pool size is largely a function of exposure risk policy. Once that policy is in place, the item pool size follows naturally (Dallas, Wang, Furter, & Luecht, 2012). And once the desired exposure rate is determined, this factor can be well controlled by the number of active panels and the number of modules contained in each panel. In some ca-MST study, the item pool size is determined as 1.5 times of the necessary number of items for the design (e.g. Wang et al. 2012), which allows for some degree of flexibility.

Ca-MST Panel Construction for Multidimensional Assessments

The framework for integrating a ca-MST test delivery model in the context of formative tests was developed recently (Luecht, 2012). In Luecht's framework, a MIRT model was used, yet simple structured items were proposed to measure each dimension in order to resolve the indeterminacy related to oblique factor structures.

To build "MIRT-sensitive" modules, Luecht's method (2012) suggested assigning items from all relevant dimensions to each module – similar to having multiple content requirements per module. If *k* traits are measured in a test, each module will incorporate *k* item sets, each of which measures one trait. The difficulty level of these item sets in a module could differ because examinees may be high on one attribute while low on the other(s). Suppose three attributes are measured, and each of them has three adaptive levels, nine item sets will be needed in a ca-MST stage (as seen in Table 1). These nine

item sets can be mixed and match to obtain 27 modules for examinees with different proficiency combinations.

The multidimensional ca-MST framework can be perceived as a combination of *k* (*k* is the number of dimensions) unidimensional ca-MST, as suggested by Luecht (2012). Table 1 represents the design for a multidimensional stage that has three levels of difficulty. It can also be perceived as a three-level stage for each of the attributes. Designing a multidimensional stage in this way can accommodate for examinees with all possible proficiency combinations while reducing the item bank demand.

Table 1. Levels of Item Set Difficulty by Trait

θ_1	θ_2	θ_3
Low	Low	Low
Medium	Medium	Medium
High	High	High

Method to Provide Reliable Subscores

When multiple traits are tackled, in order to form a test of a realistic length, the number of items measuring each trait becomes limited. To obtain reliable subscores with limited number of items, methods that regress to the group mean (such as Kelley's equation) were proposed. Under the simple structure assumption, although subscales can be estimated separately with these methods, incorporating the correlation structure in the estimation procedure using a multidimensional framework results in more precise and accurate estimation (de la Torre, 2008; de la Torre & Patz, 2005; Wang, Chen, & Cheng, 2004).

De la Torre, Song and Hong (2011) compared four methods of IRT subscoring, three of which capitalized on the inter-trait correlation, namely, multidimensional scoring, augmented scoring and higher order IRT scoring. The results suggested that these three correlation-based approaches gave highly comparable results except for extreme abilities where higher order IRT and multidimensional scoring may perform better.

Although MCMC procedure was needed for multidimensional scoring in De la Torre et al. (2011), the estimation process could have been largely simplified using empirical Bayesian method if the correlation matrix was known. In this case, the information carried by inter-correlation is still fully incorporated.

The Present Study

Although the multidimensional ca-MST framework has been constructed, the efficiency of the design and the advantages of using this framework have not been evaluated. First, we need to verify that the complexity in developing a multidimensional ca-MST scheme is worth the effort by examining its benefits. Second, this study will explore two factors – attribute correlation and item bank characteristics – that may influence the accuracy and efficiency of multidimensional ca-MST. These two factors are chosen because we are interested in the augmented information that correlated attributes could bring in. While attribute correlation is apparently an influential factor that contributes to collateral information among attributes, item bank in a large extent determines how informative a test could be by constraining item supply.

The present study aims at answering four research questions as follows:

1. Is there a benefit using multidimensional ca-MST rather than separate unidimensional ca-MSTs?
2. How does the correlation among attributes affect the accuracy and efficiency of multidimensional ca-MST?
3. How does item pool characteristic affect the accuracy and efficiency of multidimensional ca-MST?
4. Which type of panel configuration works the best under which conditions?

CHAPTER III

METHODS

This research explores the multidimensional ca-MST that measures four latent traits under the simple structure assumption. It was conducted solely by computer simulation in R. Computer simulation studies are often carried out as necessary verifications before a new testing approach can be implemented. Many ca-MST results reviewed in the previous chapter were based on computer simulations. Although computer simulations almost always fail to replicate the full picture of the reality, they are inexpensive experiments that can generally capture the trend of real testing scenario and the characteristics of new methodology.

Conditions of Study

The conditions explored in this study covered four domains: the multidimensional ca-MST panel design configurations, the item set size, the correlation among multiple measured traits, and the item pool characteristics. All levels of four variables were fully crossed to generate 108 distinct combination conditions. Under each condition, a multidimensional ca-MST process and a sequential unidimensional ca-MST process were simulated. They were replicated 20 times to obtain stable estimates. Table 2 displays the factors and levels of this simulation study.

Table 2. Simulation Factors

Factor		Levels		
ca-MST Configurations (per dimension)		"1-3"	"1-2-3"	"1-3-3"
Sub-module Size (per dimension)		3	5	8
Correlation of Traits		0.2	0.5	0.8
Item Pool Characteristics	μ_b	-1.0	0.0	
	$\boldsymbol{\mu}_a$	1.0	0.6	

Multidimensional ca-MST Configurations

In a unidimensional ca-MST design, a single latent trait is measured. For an adaptive stage, modules optimized for different ability groups can be developed by targeting different difficulty regions of the latent trait scale. For example, a two-module stage includes an easy and a difficult module that are optimized for the low and the high ability group respectively; and a 3-module stage that incorporates an easy, a medium, and a difficult module offers higher adaptation by optimizing modules for finer groups of low, medium and high proficiency.

When multiple traits are involved, the difficulty of a module and the definition of ability groups become complex because the proficiency of a person is now specified in a multidimensional space. A multidimensional module contains items that deal with different traits. In a *K* dimensional case, a module can be decomposed into *K* item sets, each of which pertains to one trait and targets at a difficulty level. However, item sets in the same module can target different difficulty levels. For example, an optimal module for a candidate who performs high on θ_1 and θ_2, but average on θ_3 and θ_4 is very likely to contain difficult item sets for the former two traits, and medium difficult sets for the later

traits. Due to the discrepancy of item set difficulties, a module does not have a difficulty level per se.

The multidimensional ca-MST framework developed by Luecht (2012) was adopted in this study. For each dimension, items targeting the same level of proficiency were grouped into an item set (also termed sub-module in this study). These item sets were then mix matched to form modules. For instance, if three adaptive alternatives were available for each trait in a four-dimensional scenario, 12 item sets were needed for module assembly, as shown in Table 3. In this case, 81 possible modules could be formed from mix matching item sets. However, some item set combinations could be dropped in situations where certain proficiency patterns were not feasible.

Table 3. Item Sets Targeting Levels of Proficiency by Trait

θ_1	θ_2	θ_3	θ_4
Low	Low	Low	Low
Medium	Medium	Medium	Medium
High	High	High	High

In this study, the ca-MST configuration conditions referred to each trait. In a "1-3" design, the first stage contained one medium difficult sub-module for each trait, and the second stage contained three sub-modules of varied difficulty level for each trait. Figure 1 and Figure 2 illustrate the multidimensional "1-2-3" design by two perspectives: separating each dimension and integrating all four dimensions. The letters "E", "M", "D", "ME" and "MD" represent Easy, Medium, Difficult, Medium Easy and Medium Difficult respectively. When seeing multiple traits separately, each dimension essentially has a

unidimensional “1-2-3” ca-MST design except that each “module” is actually an item set, or sub-module. In Figure 2, these item sets can be mix matched to create modules and each module contains four sub-modules.

Figure 1. A Unidimensional Perspetive on the “1-2-3” Multidimensional ca-MST Design

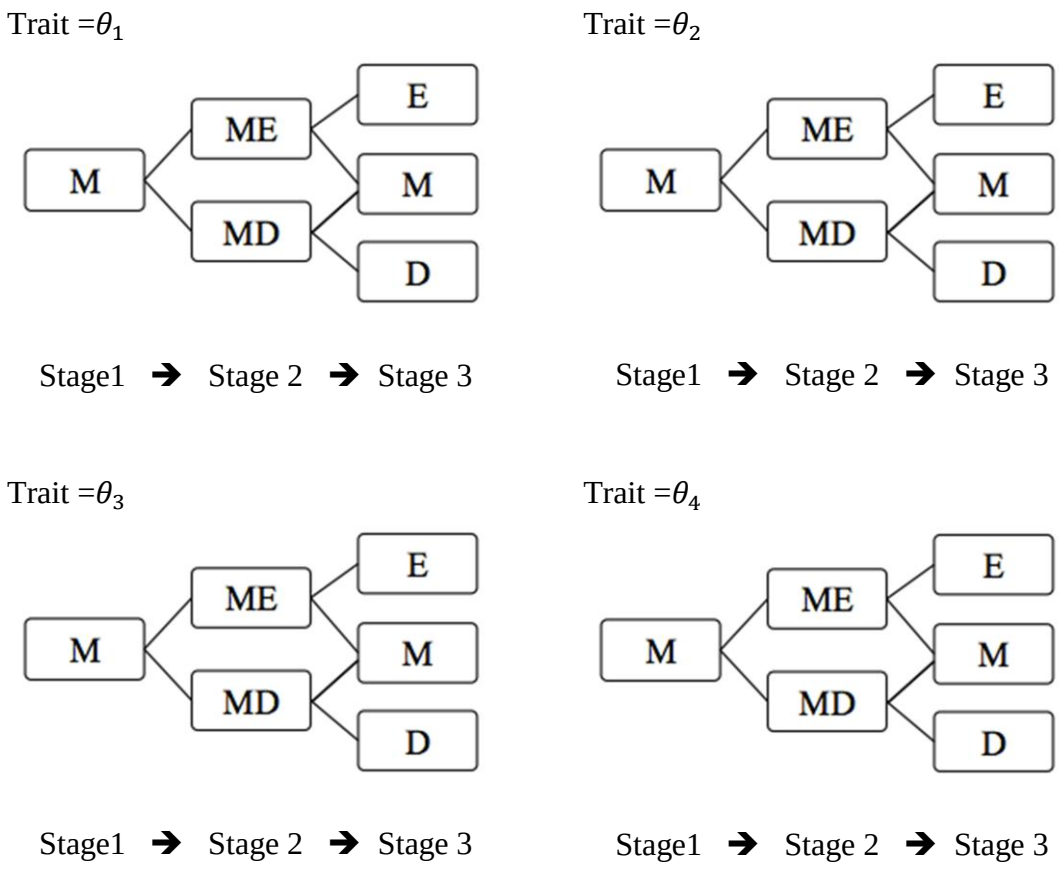
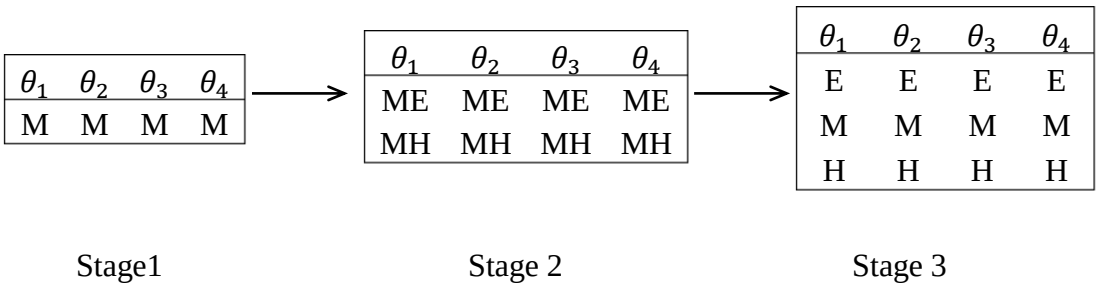


Figure 2. An Integrated Perspective on the "1-2-3" Multidimensional ca-MST Design



Sub-module Size

The number of items in a sub-module determines the length of a test module as well as the test length. In this study, the sub-module size was set to be identical for each of the four dimensions. Three levels of sub-module size (3, 5 and 8) corresponded to module sizes of 12, 20, and 32 at each stage. And the corresponding test lengths were 24, 40 and 64 items for the two-stage ca-MST design and 36, 60 and 96 items for the three-stage designs. The number of items probing each dimension ranged from 6 to 24. Three levels of item set size proposed in this study, coupled with the ca-MST configuration designs generated a reasonable range of test lengths so that an ideal length for a four-dimensional test under certain conditions might be estimated.

Test length is an important variable that influences test information and scoring precision. Generally speaking, given the same test quality, longer tests provide more information and result in more precise scoring decisions. However, test items are expensive. When multiple traits are assessed, the desired item pool size can be drastically increased, so as the cost of item bank construction. Therefore, proper test length is desirable to balance between the scoring precision and the item writing cost. On the

other hand, the compensation of information among correlated traits could reduce the demand of items. Setting a sub-module size from 3 provides an opportunity to search for economic solutions.

Correlation among Traits

Table 4 displays three levels of attribute correlation in this study[1]. Under each condition, the correlation of each pair of traits was set to be identical. Three levels of correlation coefficients represented uncorrelated, moderately correlated and highly correlated traits respectively.

Table 4. Levels of Inter-correlation

	Level 1				Level 2				Level 3			
	θ_1	θ_2	θ_3	θ_4	θ_1	θ_2	θ_3	θ_4	θ_1	θ_2	θ_3	θ_4
θ_1	1				1				1			
θ_2	0.2	1			0.5	1			0.8	1		
θ_3	0.2	0.2	1		0.5	0.5	1		0.8	0.8	1	
θ_4	0.2	0.2	0.2	1	0.5	0.5	0.5	1	0.8	0.8	0.8	1

The correlation among measured traits should affect the efficiency of multidimensional ca-MST in a large extent. Uncorrelated traits do not share information with each other and therefore, tackling multiple dimensions simultaneously in one test may not have any benefit over measuring each trait separately. On the other hand, highly

[1] We assume the true attribute correlation is known in this study. In practice, an attribute correlation matrix could be estimated from previous test scores or from the MIRT calibration with more complex method (e.g. MCMC as mentioned in De la Torre, Song and Hong, 2011).

correlated traits may actually collapse together and suggest unidimensionality. In this case, designing a multidimensional test may not even worth the effort.

In addition, inter-trait correlation influences a series of variables, including but not limited to scoring precision, item bank requirement, and the number of legitimate multidimensional routes. The former two variables relate to the amount of information shared among traits. When traits are reasonably correlated, information about one dimension can be obtained from multiple sources, which increases scoring precision. Likewise, by borrowing information from correlated traits, extra information does not have to be attained through adding items. Therefore, the demand for item bank shrinks. The third variable relates to the likelihood of ability profiles. If traits are uncorrelated, examinees can be everywhere in a multivariate space. They can be sophisticated in one skill while terrible in the others. This requires all possible item set combinations within a stage. Correlated traits, on the contrary, pose some restriction on examinees' score profiles. Diverse proficiency levels in multiple traits become less likely, and even impossible. This renders some item set combinations dispensable.

Item Pool Characteristics

Although all items were simple structured in this study, for generalization purpose the three parameter logistic (3PL) MIRT hybrid model was used. The probability of a correct response was modeled as

$$P(X_i = 1|\boldsymbol{\theta}) = c_i + \frac{1-c_i}{1+\exp[-D\boldsymbol{a}_i'(\boldsymbol{\theta}-b_i)]}, \quad (3)$$

where $\boldsymbol{\theta}$ was the vector of proficiency scores on multiple dimensions, $\boldsymbol{a}_i$ was a vector of discrimination index, b_i was the item difficulty, and the guessing parameter c_i was fixed at 0.15. For simple structured items, $\boldsymbol{a}_i$ contained a non-zero value on only one of the dimensions, which rendered this 3PL MIRT model equivalent to a unidimensional 3PL model.

Four different item pools characteristics were determined by two fully crossed two-level variables: mean item difficulty and mean item discrimination, as shown in Table 5. Locating item difficulty at -1 or 0 created conditions where item pools were either too easy or just right for the population. Generally, an item discriminative parameter *a* larger than 1.0 is considered as informative, while smaller than 0.8 is considered as not informative. In this study, item pools that had average discrimination index of 1.0 or 0.6 on every dimension indicated informative versus not informative pools.

Incorporating item difficulty as a factor relates to practical consideration. It is always easier to write items with lower difficulty. If average item difficulty does not play an essential role, more flexibility can be obtained when constructing item banks. The discrimination power of item pool determines the quality of available items. Discriminative item pools allow the assembly of informative tests, which result in better scoring decisions. Examining the main effects of item pool difficulty and discrimination, as well as their interaction can help with the design of item pool in practice.

Table 5. Item Pool Characteristics Conditions

	μ_a = (1,1,1,1)	μ_a = (.6, .6, .6, .6)
μ_b = -1	Informative Easy Bank	Uninformative Easy Bank
μ_b = 0	Informative Moderate Bank	Uninformative Moderate Bank

Data Generation

The data generation process of the study included generating proficiency scores $\boldsymbol{\theta}_i$ for test candidates and generating item pools. The proficiency profile of an examinee was a vector with a length equal to the number of measured dimensions, that was, $\boldsymbol{\theta}_i = (\theta_{1,} \theta_{2,} \theta_{3,} \theta_4)$. Under each of the 108 conditions, 20 samples of 3000 examinees were generated following a common multivariate normal distribution. Because three levels of trait correlation – low, medium, high – were involved in this study, three populations were used for sampling. They shared the same mean vector (0, 0, 0, 0), but differed in covariance matrixes as displayed in Table 4. Each population was used in 36 conditions.

Generating item pools was much more complex. To take advantage of the ca-MST mode, item pools need to be constructed corresponding to the specific ca-MST design. Factors that need to be incorporated included ca-MST configuration, sub-module size, exposure rate, and item bank characteristics. To keep the module exposure rate no more than 10%, 10 panels were assembled under each condition. Table 6 lists nine different ca-MST panel designs, the number of items needed for each module, the required number of item, and the size of reasonable item pools that allow some degree of freedom in test assembly. Under each of the nine panel design conditions, four item pools of the same size were generated according to four item pool characteristics

conditions displayed in Table 5. In doing so, a total of 9×4=36 item pools were generated.

Table 6. Item Pool Size Requirement of Different Panel Designs

Module Configuration	# of sub-module	Sub-module size	Test length	# item per panel	10 panels	Item Pool Size
"1-3"	16	3	24	48	480	720
"1-3"	16	5	40	80	800	1200
"1-3"	16	8	64	128	1280	1920
"1-2-3"	24	3	36	72	720	1080
"1-2-3"	24	5	60	120	1200	1800
"1-2-3"	24	8	96	192	1920	2880
"1-3-3"	28	3	36	84	840	1260
"1-3-3"	28	5	60	140	1400	2100
"1-3-3"	28	8	96	224	2240	3360

Since four dimensions were measured in this study, each item pool needed to contain items pertaining to all four scales. The simple structure assumption allowed us to first construct sub-pools that were composed of items measuring a single attribute. Four sub-pools were then aggregated to obtain a complete item pool. Under the condition where the mean difficulty of items in a pool μ_b = -1, each sub-pool was generated so that μ_b = -1 for k $\in$ {1, 2, 3, 4}. Item parameter b_{ki} was generated following the normal distribution with μ_k= -1, σ_k=1. μ_b = -1 conditions represented scenarios where the difficulty of item pool was not purposefully designed for ca-MST purposes. These conditions were not ideal for ca-MST because of the insufficiency in item supply for certain regions on the ability scale.

Under the conditions where μ_b = 0, the item difficulty distribution of a pool was more carefully designed so that the item pool information curve was customized for each

panel design and reflected the need of module assembly. The sub-pool size was first broken down into the sub-module level. For example, in a "1-3" design with the sub-module size of five items, each sub-pool contained 1200/4 = 300 items. These 300 items were further divided by the number of sub-modules (in this case, four) so that each sub-module had a supply of 75 candidate items. In the second step, item difficulty parameters were generated for each sub-module, which should reflect the designed module difficulty. For the previous example of "1-3" design, 75 *b* parameters were generated following N (0, 1) for stage one (medium difficult), and the same amount of items were generated for each sub-module in stage two following N (-1, 0.3), N (0, 0.3), N (1, 0.3) respectively. The same two-step process was done for all four dimensions, and the obtained difficulty parameters were aggregated to construct the item pool of a "1-3" design. Similar processes were carried out for other panel designs. The distributions used for generating items under each panel design condition are displayed in Table 7. Examples of sub-module level information curves and item pool information curves are provided in Figure 3 and Figure 4.

Table 7. Distributions for Item Difficulty Parameter Generation

Design	"1-3"	"1-2-3"	"1-3-3"
Stage 1	N (0,1)	N (0,1)	N (0,1)
Stage 2 (Sub-module 1)	N (-1, 0.3)	N (-0.5, 0.5)	N (-1, 0.5)
Stage 2 (Sub-module 2)	N (0, 0.3)	N (0.5, 0.5)	N (0, 0.5)
Stage 2 (Sub-module 3)	N (1, 0.3)	-	N (1, 0.5)
Stage 3 (Sub-module 1)	-	N (-1, 0.3)	N (-1, 0.3)
Stage 3 (Sub-module 2)	-	N (0, 0.3)	N (0, 0.3)
Stage 3 (Sub-module 3)	-	N (1, 0.3)	N (1, 0.3)

Figure 3. Examples of Sub-module Level Information Curves

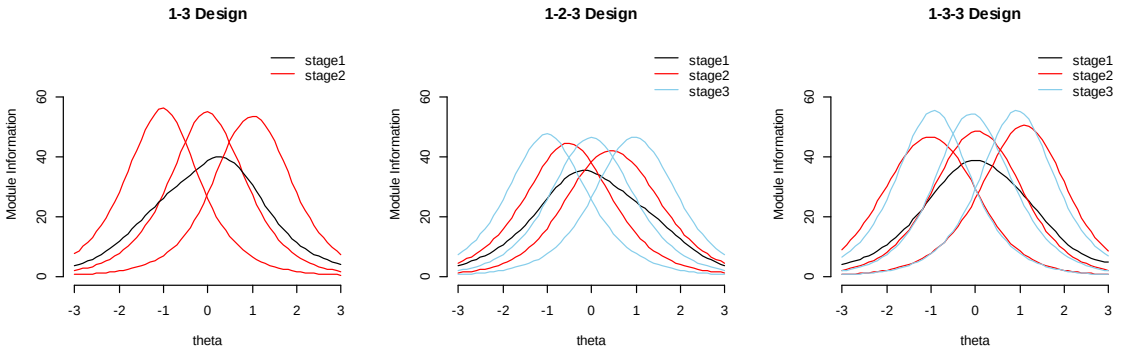
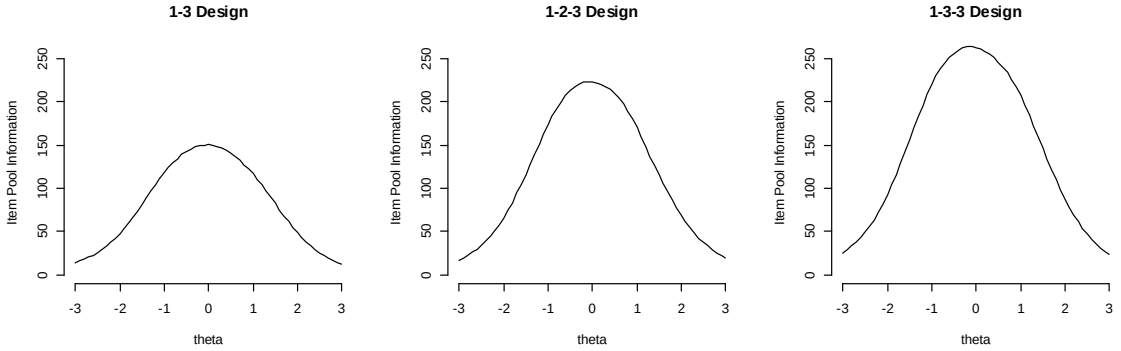


Figure 4. Examples of Item Pool Information Curves



Another factor that influences item pool quality is item discrimination parameter *a*. Similar to generating item parameter *b*, parameter *a* was first generated for each sub-pool. Under the conditions where item pools are informative, the mean item parameter μ_{ak} ($k \in \{1, 2, 3, 4\}$) was set to be 1 for each of the four sub-pools. The item discrimination parameter a_{ki} was generated following the lognormal distribution with μ_{ak} = 1. Because the mean of a lognormal distribution is $e^{\mu+\sigma^2/2}$, where μ and σ^2 are the mean and the variance of the corresponding normal distribution respectively, $\mu+\sigma^2/2=0$ satisfies $e^{\mu+\sigma^2/2}=1$. For μ_a = 1, the population from which item parameters were

generated was a lognormal distribution with corresponding normal distribution N (-1/32, 0.25).

Under the conditions of uninformative item pools, the mean item parameter μ_{ak} ($k \in \{1, 2, 3, 4\}$) was set to be 0.6 for each of the four sub-pools. The item discrimination parameters in each sub-pool were generated following a lognormal distribution likewise. Following the same calculation process, we used the lognormal distribution with the corresponding normal distribution N (ln(0.6)-1/32, 0.25) so that μ_{ak} = 0.6.

To verify the efficiency of employing multidimensional ca-MST and compare different ca-MST designs, a fixed form was simulated as the baseline condition. Because adaptive tests were expected to achieve the same scoring accuracy as traditional paper-and-pencil tests with fewer items, we determined the test length of the fixed form to be 96 items – the length of the longest ca-MST design used in this study. These 96 items were evenly distributed to measure four dimensions. Item parameters for each dimension were generated separately, yet following the same process. 24 item difficulty parameters for each dimension were generated following the normal distribution N (0, 1), and the item discrimination parameters were generated following the lognormal distribution with corresponding normal distribution N (-1/32, 0.25).

After generating item parameters, a random sample of examinees was used to generate corresponding responses to all items following a hybrid 3PL IRT model (c=0.15). These responses were then calibrated to obtain estimated item parameters for item pools.

Test Assembly

The test assembly dealt with panel assembly for multidimensional ca-MST as well as separate unidimensional ca-MSTs. The bottom-up assembly strategy was employed. The same process and algorithm applied to the sub-module assembly in the multidimensional conditions and the module assembly in the unidimensional conditions. For this reason, the assembled sub-modules of a multidimensional ca-MST were used as modules that pertained to the same attribute in the corresponding sequential unidimensional ca-MST. The multidimensional ca-MST needed one more step that assembled the sub-modules into a four-dimensional module, which happened during the test procedure. Under each condition, there was no overlap modules or items in the 10 multidimensional ca-MST panels (and it was the same with the 10 sequential unidimensional ca-MSTs).

Test Information Targets

Because the bottom-up strategy was used in test assembly, test information targets were specified in the sub-module level of the multidimensional ca-MST designs. The approximate maximum information (AMI) method (Luecht, 2000) was used to determine target TIFs. The 5-step AMI method was as follows.

1. Located a particular point on the θ scale where the desired TIF peaked. In the current study, one particular point was predetermined for each of the sub-module. For example, a moderate difficult sub-module would peak at θ=0.

2. For each item in the sub-pool pertaining to a specific attribute, computed the item information at the specified point (e.g. θ=0).
3. Sorted the sub-pool in descending order by the computed item information value. This was done for each of the sub-pools respectively.
4. Given the sub-module size n, and the number of replicated sub-modules m, the most informative n×m items at the previously determined locate were chosen.
5. For these selected $n \times m$ items, computed the sum of item information at each of the selected ability points, θ_k, k=1, …, K. The selected ability points generally cover a reasonable range of the ability scale, e.g. -3 to 3 at the increment of 0.3. Divided the aggregated information at each selected point by m, the TIF was obtained. That is,

$$\mathrm{TIF}(\theta_k) = \frac{\sum_{i=1}^{n \cdot m} I_i(\theta_k)}{m}. \quad (4)$$

Following the AMI method, 10 (because 10 panels need to be assembled) non-overlapping sub-modules that maximized information at a certain points were constructed and the 10 TIFs were averaged to obtain a provisional target TIF. Table 8 specifies the points where sub-module information was maximized. To represent increasing adaptation as stages move on, the variance of b parameters should decrease so that the later sub-module focuses on narrower region on the ability scale. Bearing this in consideration, when there were more than one sub-module of a panel maximize information at the same point (e.g. in the "1-3" design, two sub-modules maximized their

information at $\theta_k = 0$), the backward assembly strategy was used. That is, sub-modules in the later stages were assembled with priority. Generally, the most informative items have difficult parameters close to the post point. Therefore, sub-modules constructed with priority had smaller variance in *b* parameters than those constructed later.

Table 8. Points where Sub-module Information was Maximized

Design	"1-2"	"1-3"	"1-2-3"	"1-3-3"
Stage 1	$\theta_k = 0$	$\theta_k = 0$	$\theta_k = 0$	$\theta_k = 0$
Stage 2	$\theta_k = (-1, 1)$	$\theta_k = (-1, 0, 1)$	$\theta_k = (-0.5, 0.5)$	$\theta_k = (-1, 0, 1)$
Stage 3	–	–	$\theta_k = (-1, 0, 1)$	$\theta_k = (-1, 0, 1)$

Assembly algorithm

The assembly of sub-modules followed the normalized weighted absolute deviation heuristic (NWADH) (Luecht, 1998). Suppose N items were need in a sub-module, Luecht's heuristic found the *j*th item by taking two steps: 1) divided the remaining difference between the target values and current values of the information function by the remaining N-(*j*-1) items. Because θ was a continuous variable, it was not practical in test construction to consider all values of θ. Instead, discrete points on the θ scale were selected to represent θ over a reasonable range of values, and denoted as θ_q. In this study, 31 equidistant quadrature points from -3 to 3 were used. Let $T(\theta_q)$ denoted the target information function at the point of θ_q, the first step yielded objective function

$$\left[T(\theta_q) - \sum_{i=1}^{j-1} I_j(\theta_q)\right]/(n - j + 1). \quad (5)$$

2) Selected the item with an information function that matched the above quantity best over all values θ_q. In equation, this selected item should maximize

$$e_i = 1 - \frac{\sum_{q=1}^{Q} d_{iq}}{\sum_{i \epsilon R_{j-1}} \sum_{q=1}^{Q} d_{iq}} \quad , \tag{6}$$

Where R_{j-1} was the remaining item pool after selecting j-1 items,

$$d_{iq} = \left| \left[\frac{T(\theta_q) - \sum_{i=1}^{j-1} I_j(\theta_q)}{(n-j+1)} \right] - I_j(\theta_q) \right| . \tag{7}$$

Density weights could be incorporated into the summation over the Q quadrature values of d_{iq}. However, in this study we did not apply any weights.

Test Process and Scoring

In simulating test process, each examinee was randomly assigned with one of the ten multidimensional ca-MST panels. Once the multidimensional ca-MST was completed, four unidimensional ca-MST that shared the same sub-modules with the tested multidimensional counterpart were given to the examinee. Scores of each dimension were recorded under both scenarios.

In the unidimensional ca-MST process, the 3PL IRT hybrid scoring model with expected a prior (EAP) estimation method was used for both the routing and the final scoring. The estimate of θ on each dimension was obtained by

$$\hat{\theta} = \sum P(\theta|X)\theta, \tag{8}$$

where X is the item responses of an examinee, θ is the selected quadrature points on the ability scale, and $P(\theta|X)$ is the probability of an ability score given an examinee's item responses, which can be calculated by

$$P(\theta|X) = \frac{L(X|\theta)P(\theta)}{\sum L(X|\theta)P(\theta)}. \quad (9)$$

$L(X|\theta)$ is the likelihood function of getting a specific response vector X for I items given an ability score θ, and it is calculated as

$$L(X|\theta) = \prod_{i=1}^{I}(P(X_i = 1|\theta, a, b, c))^{X_i}(1 - (X_i = 1|\theta, a, b, c))^{1-X_i}. \quad (10)$$

Once a panel was assembled, the intersection of TIFs of the same stage could be determined as cut points, and used for routing examinees to a module that provides maximum information. Meanwhile, the MIRT model was used for final score decision so that the differences between the unidiemsnional ca-MST designs and their multidimensional ca-MST counterpart would not attribute to the scoring method. The EAP estimation method, again, was used for estimating the MIRT model. The estimated ability on the k^{th} attribute is

$$\hat{\theta}_k = \sum P(\theta_k|X, \Sigma)\theta_k, \quad (11)$$

where X is the item responses of an examinee, θ_k is the selected quadrature points on the k^{th} attribute, and Σ is the correlation matrix among K attributes. As a multidimensional extension of Equation 9, the probability of an ability vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ given item responses and an attribute correlation matrix is

$$P(\boldsymbol{\theta} \mid \mathrm{X}, \Sigma) = \frac{L(X|\boldsymbol{\theta},\boldsymbol{\Sigma})P(\boldsymbol{\theta})}{\sum L(X|\boldsymbol{\theta},\boldsymbol{\Sigma})P(\boldsymbol{\theta})}. \quad (12)$$

The likelihood function can be calculated as the product of the item response probability given an ability vector,

$$L(X|\boldsymbol{\theta},\boldsymbol{\Sigma}) = \prod_{i=1}^{I}(P(X_i = 1|\boldsymbol{\theta},\boldsymbol{a},b,c))^{X_i}(1-(X_i = 1|\boldsymbol{\theta},\boldsymbol{a},b,c))^{1-X_i}. \quad (13)$$

The ability vector $\boldsymbol{\theta}$**~MVN(0, Σ)** , and the density on each k-dimensional quadrature point $P(\boldsymbol{\theta})$ can be determined according to the distribution. The density on the k^{th} scale $P(\theta_k \mid X,\Sigma)$ is obtained by integrating $P(\boldsymbol{\theta} \mid X,\Sigma)$ across the other k-1 scales. Given the chosen quadrature points, the integration can be estimated by summing across k-1 scales at the quadrature points of each attribute. Take a three-dimensional case for example, we suppose that the quadrature points for each scale are chosen from -3 to 3 at an increment of 0.5 and thus have 13 points at each scale, 13^3=2197 points in total. For Scale 1, the density at the quadrature point θ_1=0.5 is

$$P(\theta_1 = 0.5|X,\Sigma) = \sum_{\theta_3=-3}^{3}\sum_{\theta_2=-3}^{3} P(\boldsymbol{\theta} = (0.5,\theta_2,\theta_3)|X,\Sigma). \quad (14)$$

In the multidimensional ca-MST process, the 3PL MIRT hybrid model was used for scoring. When routing within a panel, the sub-modules were mixed with all possible combination to create m^4 modules, where m was the number of difficulty levels defined for each dimension. In a multidimensional space, test information at a provisional point becomes a matrix. Combined with the tested module(s), the module that yielded largest determinant of test information matrix at provisional ability location was chosen. In our

study where items were simple structured, this routing method was equivalent to choosing a sub-module that maximized information at the provisional theta point for each skill. We used the later alternative so that modules did not need to be preassembled, which avoided the effort of assembling unrealistic modules. The MIRT EAP scoring method was used for routing as well as final scoring decision.

For the baseline condition, the fixed form were given to three examinee samples of size 3000, sampling from four-dimensional multivariate normal distributions with marginal distribution N(0,1) and three correlation matrix as shown in Table 4. Both IRT and MIRT models were employed for scoring.

Evaluation of Results

The analysis of results mainly focused on the precision of scoring decisions, which was evaluated by the mean bias and the root mean square error (RMSE) of the estimated scores. Mean bias was calculated as

$$\bar{e} = \frac{\sum_{j=1}^{N}(\hat{\theta}-\theta)}{N}, \tag{15}$$

where N was the number of examinees, $\hat{\theta}$ was the estimated ability score, and θ was the true ability score. And RMSE was obtained using Equation 16.

$$\text{RMSE} = \sqrt{\frac{\sum_{j=1}^{N}(\hat{\theta}-\theta)^2}{N}}. \tag{16}$$

The first analysis dealt with the question whether multidimensional ca-MST had any benefit over separate unidimensional ca-MSTs. The mean bias and RMSE were calculated for both the unidimensional and the multidimensional scenario under each of the 108 conditions. Because the main difference between these two scenarios fell in the routing where compensation among traits may or may not happen, we examined these differences when item bank varies and inter-trait correlation differs.

The other three research questions focused on the comparisons within the multidimensional ca-MST scenarios. Question 2 and Question 3 were answered by looking at the main effect of the attribute correlation and the item pool characteristics. To answer Question 4, the interaction between attribute correlation and item pool characteristics were examined under each of the 9 panel designs. The fix-form condition was used as the baseline in comparison.

CHAPTER IV

RESULTS

The results of analysis are presented in five sections. The first section describes the simulated item pools, the constructed TIF, and the quality of assembled multidimensional ca-MST panels. Section two to five summarize the simulation results to answer the four research questions respectively.

Multidimensional ca-MST Panel Assembly

Calibrated Item Pool

The simulated item parameters, along with an examinee sample of 3000 were used to generate item responses to each item. Because of the superimposed simple structure assumption, each sub-pool was calibrated separately using hybrid 3PL IRT model (c=0.15). The discrimination and difficulty parameters of the sub-pools were summarized in Table 9 and Table 10 respectively.

Table 9. Mean and Standard Deviation of Discrimination Parameters

Condition	Mean				sd			
Design_sub-module size_a_b	a1	a2	a3	a4	a1	a2	a3	a4
1-3_3_informative_easy	1.03	1.05	1.03	1.01	.26	.26	.24	.27
1-3_3_informative_moderate	.98	1.03	1.01	1.02	.25	.29	.29	.27
1-3_3_uninformative_easy	.59	.59	.60	.61	.15	.14	.16	.15
1-3_3_uninformative_moderate	.56	.60	.61	.60	.13	.16	.15	.15
1-3_5_informative_easy	.99	1.01	1.01	1.02	.27	.25	.27	.26
1-3_5_informative_moderate	.97	1.02	1.01	1.02	.25	.26	.25	.27
1-3_5_uninformative_easy	.58	.61	.62	.60	.15	.15	.16	.16
1-3_5_uninformative_moderate	.61	.62	.62	.61	.15	.15	.16	.15
1-3_8_informative_easy	.99	1.05	1.00	1.05	.24	.25	.26	.25
1-3_8_informative_moderate	1.02	1.02	1.04	1.03	.27	.26	.25	.27
1-3_8_uninformative_easy	.58	.62	.61	.62	.15	.16	.15	.16
1-3_8_uninformative_moderate	.59	.62	.62	.62	.15	.15	.16	.16
1-2-3_3_informative_easy	.98	1.02	1.01	1.05	.24	.24	.23	.26
1-2-3_3_informative_moderate	1.01	1.04	1.02	1.00	.27	.28	.26	.25
1-2-3_3_uninformative_easy	.59	.61	.63	.60	.14	.14	.16	.16
1-2-3_3_uninformative_moderate	.58	.62	.62	.61	.15	.16	.15	.15
1-2-3_5_informative_easy	.98	1.04	1.03	1.03	.25	.29	.30	.27
1-2-3_5_informative_moderate	1.01	1.03	1.05	1.03	.26	.27	.26	.27
1-2-3_5_uninformative_easy	.59	.61	.61	.61	.15	.15	.16	.15
1-2-3_5_uninformative_moderate	.59	.63	.63	.62	.15	.15	.16	.15
1-2-3_8_informative_easy	.99	1.03	1.03	1.05	.26	.26	.26	.28
1-2-3_8_informative_moderate	.99	1.06	1.04	1.02	.25	.27	.27	.25
1-2-3_8_uninformative_easy	.61	.61	.62	.61	.16	.16	.17	.16
1-2-3_8_uninformative_moderate	.60	.63	.61	.63	.16	.17	.15	.16
1-3-3_3_informative_easy	.97	1.05	1.01	1.02	.24	.29	.26	.24
1-3-3_3_informative_moderate	1.00	1.02	1.00	1.03	.26	.26	.28	.24
1-3-3_3_uninformative_easy	.60	.63	.60	.60	.15	.16	.16	.15
1-3-3_3_uninformative_moderate	.60	.62	.61	.63	.17	.17	.16	.14
1-3-3_5_informative_easy	1.00	1.02	1.04	1.02	.25	.25	.29	.25
1-3-3_5_informative_moderate	.99	1.04	1.05	1.04	.24	.27	.29	.27
1-3-3_5_uninformative_easy	.61	.62	.61	.61	.15	.16	.15	.15
1-3-3_5_uninformative_moderate	.59	.61	.62	.61	.14	.15	.15	.15
1-3-3_8_informative_easy	1.01	1.03	1.03	1.02	.26	.25	.28	.25
1-3-3_8_informative_moderate	1.00	1.04	1.04	1.04	.24	.26	.26	.27
1-3-3_8_uninformative_easy	.60	.61	.62	.63	.15	.15	.16	.16
1-3-3_8_uninformative_moderate	.60	.62	.62	.61	.16	.16	.16	.16

Table 10. Mean and Standard Deviation of Difficulty Parameters

Condition	Mean				sd			
Design_sub-module size_a_b	b1	b2	b3	b4	b1	b2	b3	b4
1-3_3_informative_easy	-1.00	-1.02	-.92	-.91	1.00	.95	.95	.88
1-3_3_informative_moderate	.05	-.07	.03	.03	.96	.92	.90	.88
1-3_3_uninformative_easy	-.92	-.95	-1.07	-.94	1.13	1.09	1.08	.94
1-3_3_uninformative_moderate	.06	.00	.04	.01	.94	.96	.86	.93
1-3_5_informative_easy	-1.07	-.98	-.98	-1.04	1.05	.96	.97	.95
1-3_5_informative_moderate	.07	-.06	.00	-.02	.93	.85	.90	.89
1-3_5_uninformative_easy	-.91	-.92	-1.02	-.98	.99	.94	.93	.97
1-3_5_uninformative_moderate	.00	-.04	-.04	.02	.92	.84	.86	.92
1-3_8_informative_easy	-1.06	-.97	-1.04	-1.10	1.01	.92	.95	.95
1-3_8_informative_moderate	.04	.02	-.03	.00	.93	.87	.87	.88
1-3_8_uninformative_easy	-.96	-.97	-.96	-1.08	1.02	.98	1.03	.96
1-3_8_uninformative_moderate	.07	-.02	-.02	-.02	.94	.84	.90	.90
1-2-3_3_informative_easy	-.97	-1.04	-.96	-1.03	1.02	1.02	.99	1.01
1-2-3_3_informative_moderate	.00	-.03	-.03	-.01	.83	.84	.81	.83
1-2-3_3_uninformative_easy	-.97	-.97	-.95	-1.03	1.07	.95	1.00	.98
1-2-3_3_uninformative_moderate	.00	.01	-.01	-.06	.89	.82	.84	.84
1-2-3_5_informative_easy	-.96	-.90	-.90	-.97	1.00	.97	.92	.94
1-2-3_5_informative_moderate	.02	-.04	-.02	.01	.85	.81	.79	.81
1-2-3_5_uninformative_easy	-.95	-1.01	-.99	-1.01	.97	.99	1.01	1.01
1-2-3_5_uninformative_moderate	.02	-.03	-.02	-.04	.84	.84	.80	.82
1-2-3_8_informative_easy	-.94	-1.02	-.94	-.98	1.00	.92	.98	.90
1-2-3_8_informative_moderate	-.02	-.03	.00	-.01	.82	.81	.79	.81
1-2-3_8_uninformative_easy	-.93	-1.02	-.95	-.99	.96	.94	.96	.93
1-2-3_8_uninformative_moderate	-.01	-.02	-.01	-.05	.87	.80	.85	.83
1-3-3_3_informative_easy	-1.02	-.91	-.91	-.99	.97	.89	1.01	.91
1-3-3_3_informative_moderate	.05	-.03	-.01	-.01	.93	.88	.91	.86
1-3-3_3_uninformative_easy	-.94	-.97	-.97	-.91	1.04	.95	.93	.96
1-3-3_3_uninformative_moderate	-.01	.00	.01	.01	.99	.90	.92	.95
1-3-3_5_informative_easy	-1.01	-.98	-.94	-1.05	1.01	.94	.96	.92
1-3-3_5_informative_moderate	.03	-.03	-.03	-.03	.91	.88	.87	.89
1-3-3_5_uninformative_easy	-.96	-1.04	-.98	-.98	1.03	.98	1.01	.99
1-3-3_5_uninformative_moderate	.00	.00	.01	-.03	.91	.93	.90	.88
1-3-3_8_informative_easy	-.98	-1.01	-.99	-.98	1.05	.98	.95	.95
1-3-3_8_informative_moderate	.01	.00	.01	-.04	.93	.91	.87	.91
1-3-3_8_uninformative_easy	-.97	-1.06	-1.03	-1.01	1.04	.94	1.01	1.00
1-3-3_8_uninformative_moderate	.01	-.04	-.02	-.04	.93	.92	.91	.89

As shown in Table 9, the mean of discrimination parameters is 1.0 or approximate to 1.0 for informative item pools, and 0.6 or approximate to 0.6 for uninformative item pools. The standard deviation of the *a* parameters is about 0.25 for the informative pools and 0.15 for the uninformative pools. The mean of difficulty parameter is close to 0.0 for the moderate difficult pools and -1.0 for the easy pools. The standard deviation of b parameter ranges from 0.88 to 1.13 for the easy pools, while that for the moderate pool is comparatively smaller, ranging from 0.79 to 0.96, as suggested in Table 10.

Target Information Function

Using the AMI method, TIFs were found for each sub-module at each stage. Figure 5 through Figure 7 present three examples of TIFs for informative easy pools[2]. The TIFs constructed under uninformative conditions were of very similar patterns, except for flatter curves.

[2] The corresponding colors used for TIFs in each design are: 1 -3(blue – black, red, green); 1-2-3 (pink – blue, light blue – black, red, green) ; 1-3-3 (yellow- blue, light blue, pink – black, red, green).

Figure 5. Sub-module TIFs for the 1-3 Design with Easy Pools

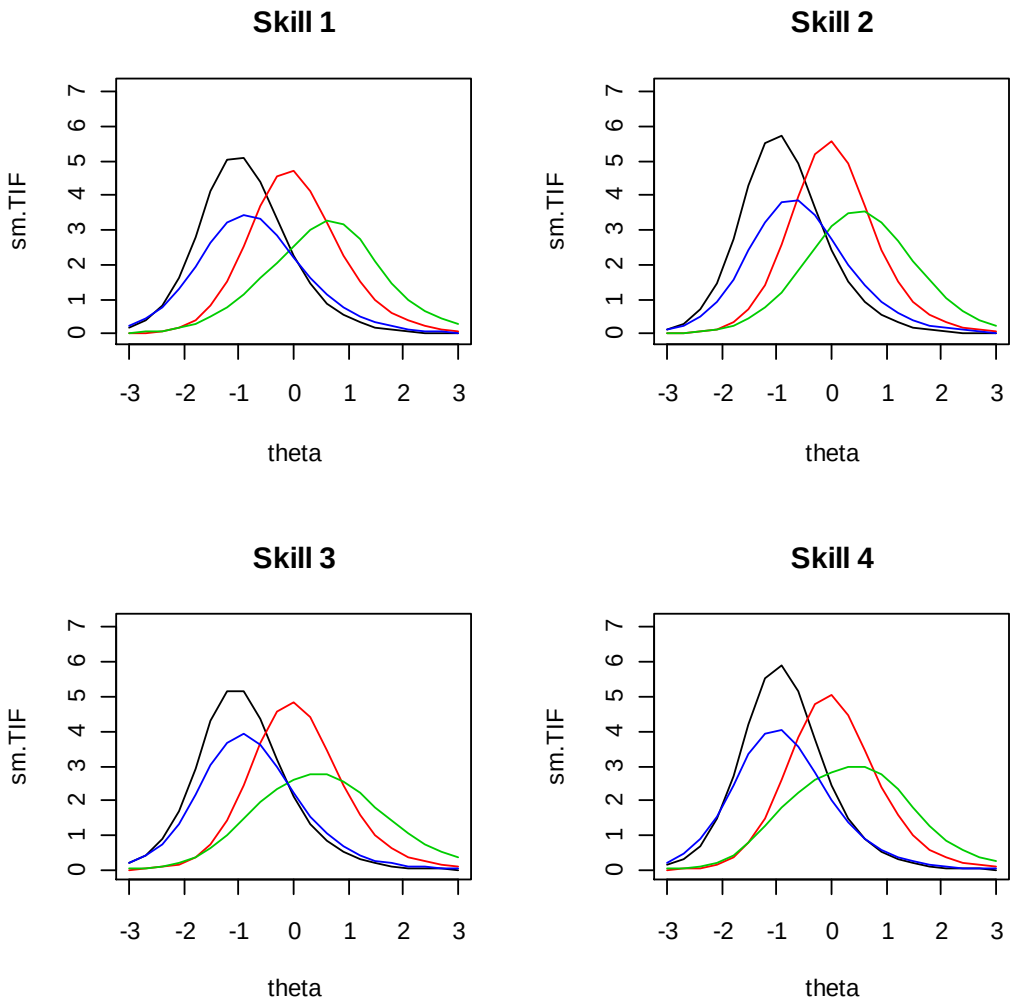


Figure 6. Sub-module TIFs for the 1-2-3 Design with Easy Pools

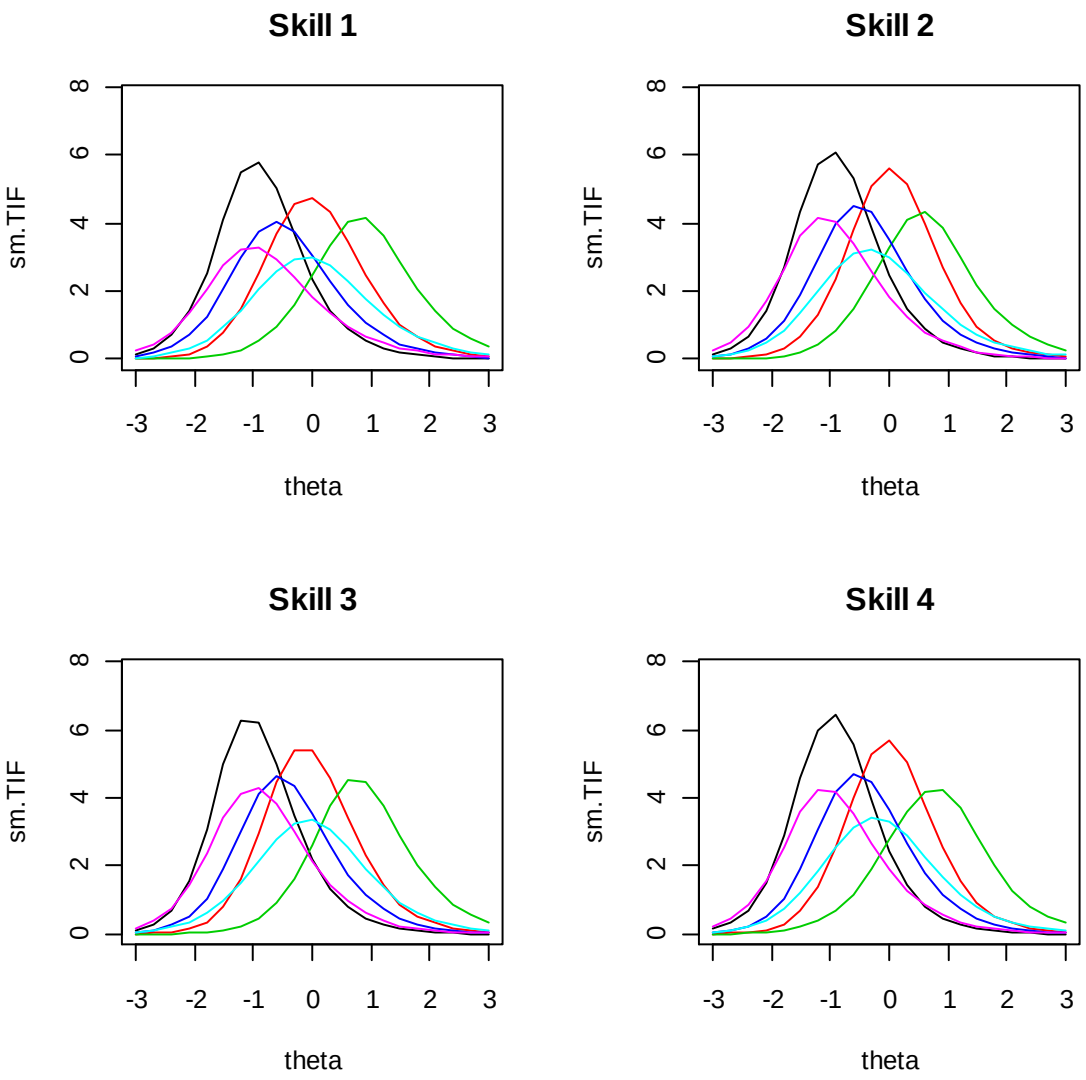


Figure 7. Sub-module TIFs for the 1-3-3 Design with Easy Pool

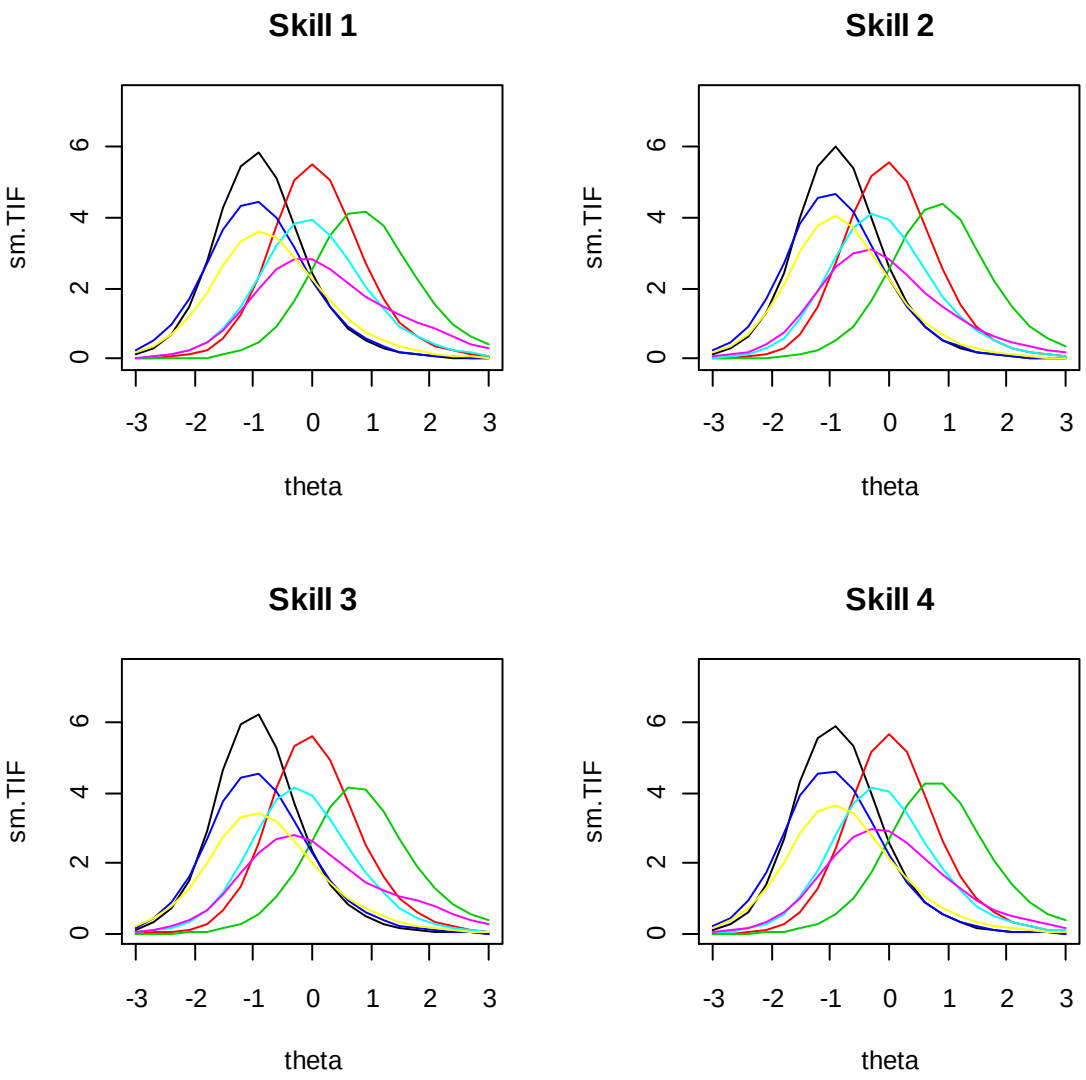


Figure 5 to Figure 7 suggest that with an easy item pool, the constructions of TIFs are limited by the lack of items at the higher half of the scale. The assembled tests, therefore, would be informative at the lower end but uninformative for examinees with moderate to high proficiency. From the unidimensional perspective and look at the TIFs for each skill respectively, the reliability of assessment would be questioned when the test

is given to people with higher performance levels. Because the difficulty levels of the alternatives in the same stage could be very close, very limited adaptation is available.

Three examples of constructed TIFs for moderate difficult pools are displayed in Figure 8 to 10. These figures represent ideal TIFs for a ca-MST: 1) each TIF locates at the designed difficulty region of the scale and maximizes its information at the desired post point; 2) later stage(s) has sharper TIFs, indicating sub-modules with items that are more similar in difficulty.

Figure 8. Sub-module TIFs for the 1-3 Design with Moderate Difficult Pool

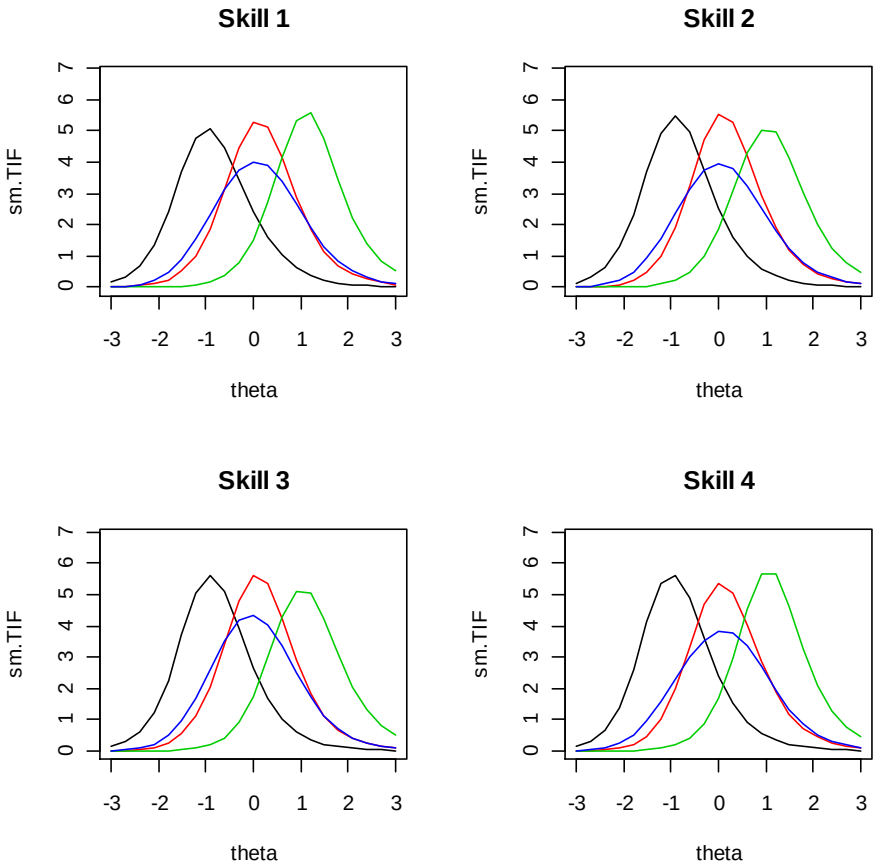


Figure 9. Sub-module TIFs for the 1-2-3 Design with Moderate Difficult

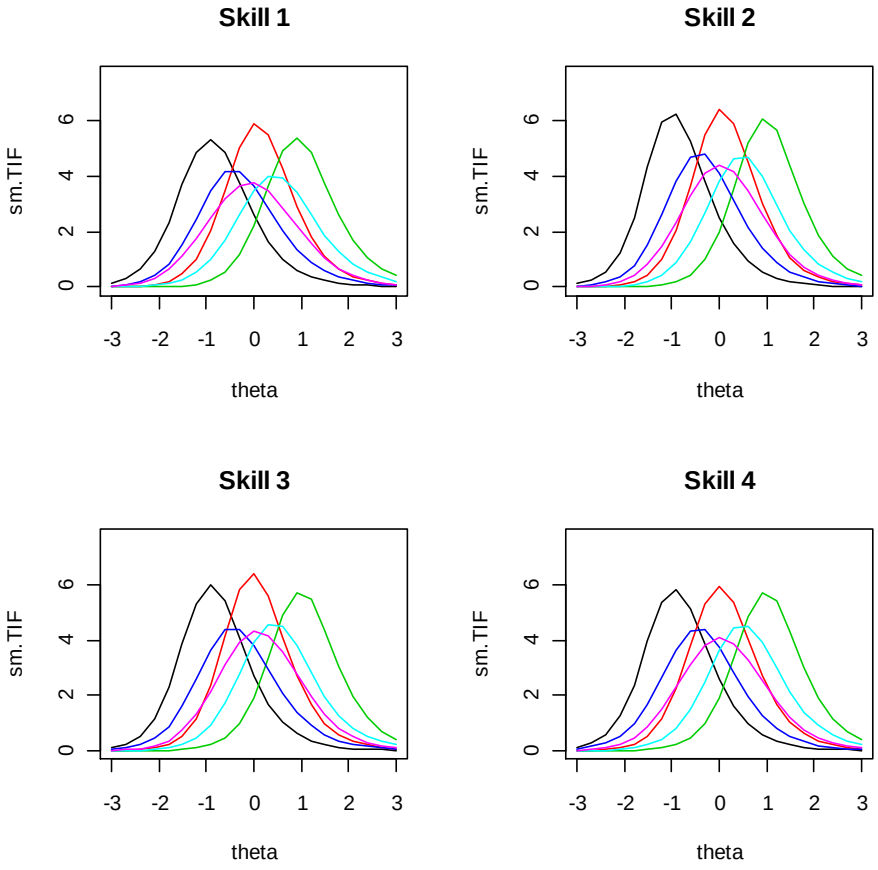
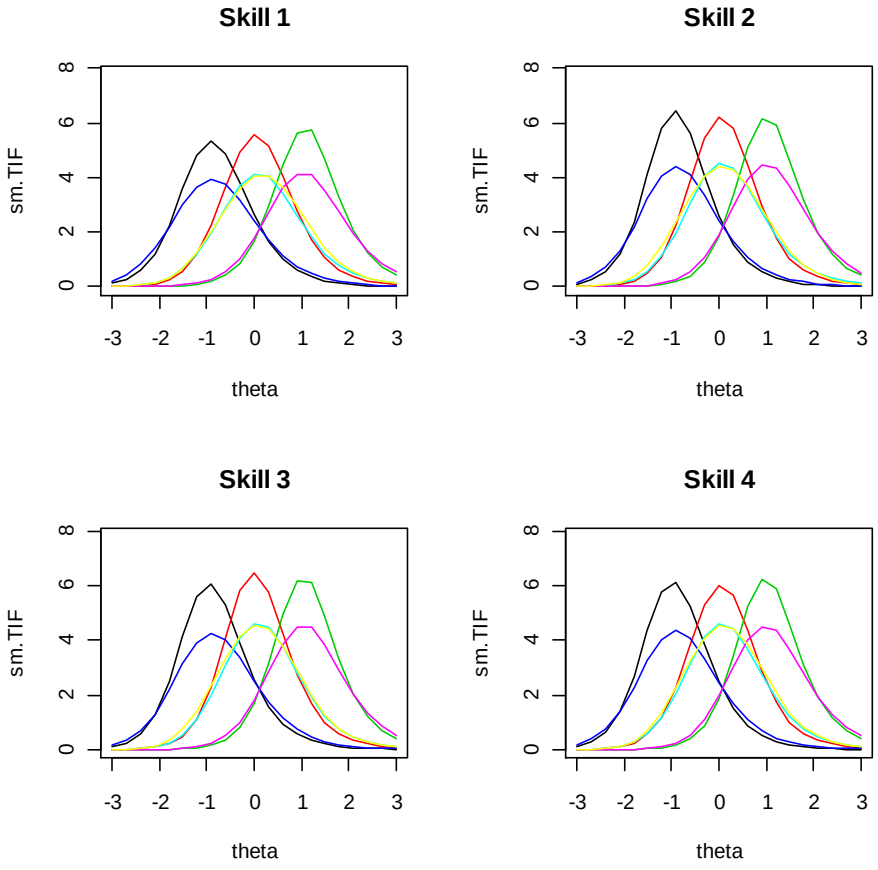


Figure 10. Sub-module TIFs for the 1-3-3 Design with Moderate Difficult Pool



Panel Assembly Quality

To evaluate the quality of panel assembly, the sub-module TIFs of the same trait were aggregated to obtain a "sub-panel"[3] TIF. This "sub-panel" TIF was then compared with the empirical "sub-panel" information curves after ten replicated "sub-panels" were assembled.

Figure 11 to 13 display some examples of the "sub-panel" TIF versus empirical curves for a single skill. Again, all figures were plotted for the informative pool

[3] A "sub-panel" contains all item sets of the same trait, as seen in Figure 1. Although it is very similar to a unidimensional ca-MST panel, it does not indicate any administration unit.

conditions because the uninformative conditions yield similar patterns except for flatter curves. The red curves in the figures are "sub-panel" TIFs, and the black curves are ten replications of empirical information curves. As shown in Figure 11 to 13, the empirical information curves are very close to the TIFs, suggesting successful "sub-panel" assembly. Similar results were found for each of the measured skills, and the quality of test assembly was not affected by test length.

Figure 11. "Sub-panel" Information Curves for 1-3 Design

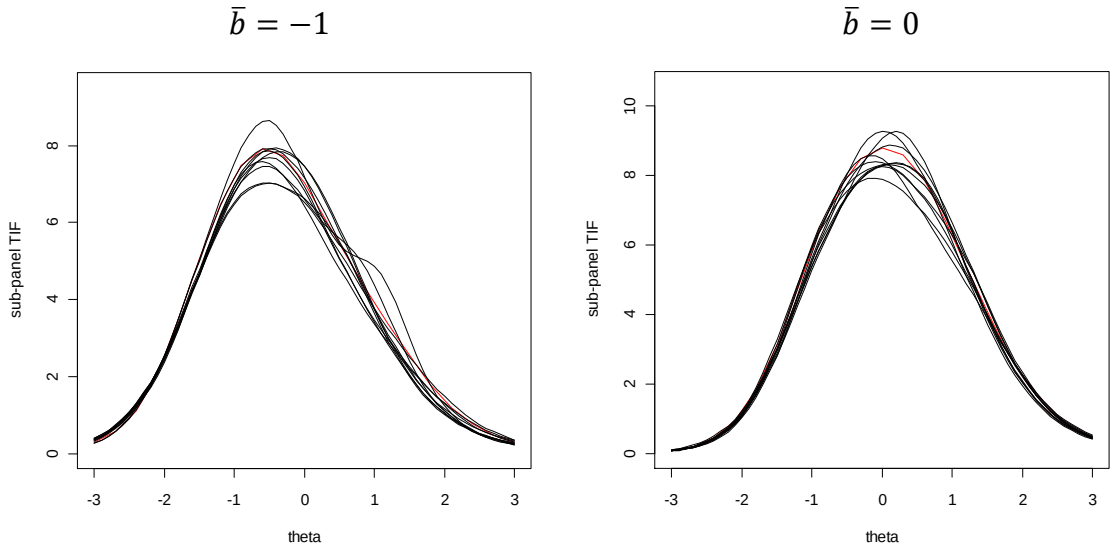


Figure 12. “Sub-panel” Information Curves for 1-2-3 Design

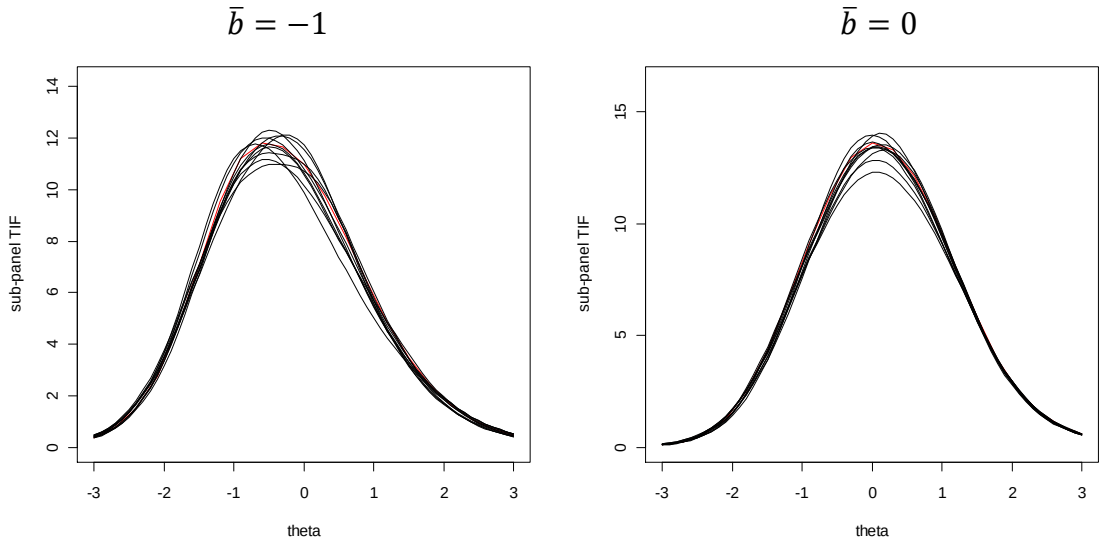
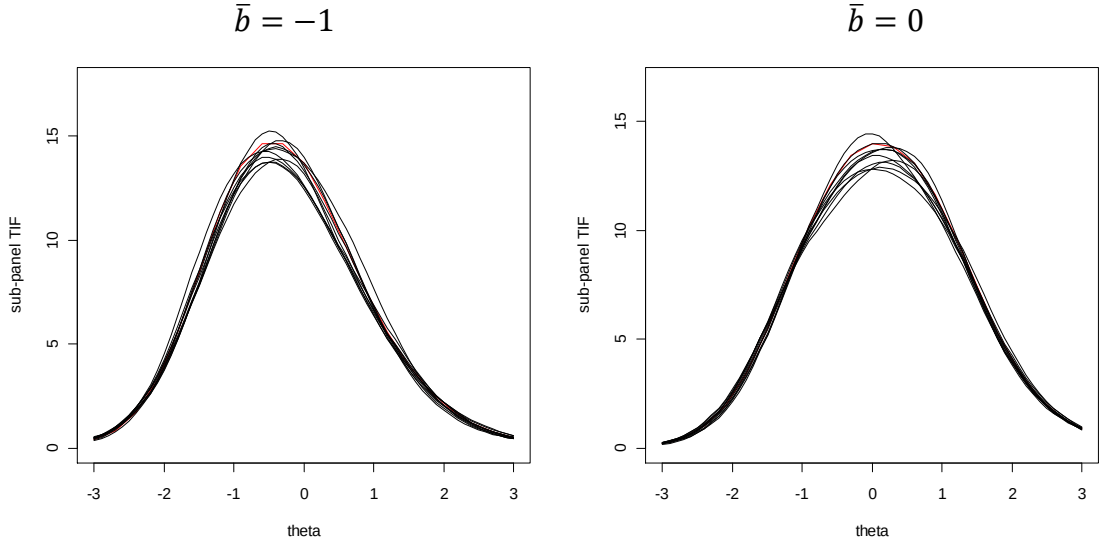


Figure 13. “Sub-panel” Information Curves for 1-3-3 Design



Unidimensional versus Multidimensional ca-MST

The results of multidimensional ca-MST were compared to its unidimensional ca-MST counterpart, which contained four separate panels and employed both IRT and MIRT scoring methods. The bias, RMSE, and residual correlation under each condition were obtained by averaging across 20 replications.

The mean bias of scoring results was displayed in Table 11. For simplification purpose, biases were averaged across three sub-module size conditions. As suggested in Table 11, it is evidential that item pool characteristics affect the magnitude of bias, which will be discussed in more details in the later section. In addition, employing MIRT scoring under the unidimensional ca-MST scenario yielded slightly larger bias. A repeated measure analysis suggested strong evidence that significant differences existed between multidimensional and unidimensional MST (both IRT and MIRT scoring) in respect of scoring bias ($F=20.224$, $df=2$, $p<0.001$). Post-hoc multiple comparisons suggested no significant difference between multidimensional ca-MST and unidimensional IRT results ($p=0.401$), but unidimensional MIRT scoring was more biased than the former two methods ($p<0.001$ for both). In spite of the statistically significant difference, the magnitude of bias under the unidimensional MIRT scoring condition was still very small, and would not be considered as a concern in score reporting.

Table 11. Mean Bias of Scoring Results with Multidimensional and Unidimensional ca-MST

r	Configuration	a	b	Multi.MST	Uni.IRT	Uni.MIRT
0.2	1-3	1	-1	.010	.009	.010
			0	-.001	.000	.000
		.6	-1	.005	.006	.007
			0	.005	.005	.006
	1-2-3	1	-1	.010	.010	.011
			0	.001	.002	.002
		.6	-1	.010	.010	.011
			0	.004	.005	.005
	1-3-3	1	-1	.010	.011	.011
			0	.000	.001	.001
		.6	-1	.007	.007	.008
			0	.002	.004	.004
0.5	1-3	1	-1	.009	.007	.010
			0	.001	.001	.002
		.6	-1	.007	.005	.007
			0	.008	.006	.007
	1-2-3	1	-1	.010	.010	.011
			0	.000	.002	.002
		.6	-1	.010	.008	.010
			0	.005	.005	.006
	1-3-3	1	-1	.011	.011	.013
			0	.000	.001	.001
		.6	-1	.008	.008	.011
			0	.002	.002	.003
0.8	1-3	1	-1	.012	.008	.011
			0	.000	.002	.003
		.6	-1	.008	.006	.008
			0	.006	.005	.007
	1-2-3	1	-1	.011	.009	.011
			0	.000	.001	.001
		.6	-1	.014	.010	.013
			0	.005	.006	.006
	1-3-3	1	-1	.010	.009	.012
			0	-.001	.001	.001
		.6	-1	.011	.009	.012
			0	.002	.002	.004
Mean				.006	.006	.007
Sd				.004	.003	.004

Table 12 shows the mean RMSE of scoring results. Similarly, the mean RMSE is affected by item pool characteristics – lower RMSE is accompanied with higher discrimination parameters and moderate item pool difficulty. A repeated measure analysis suggested strong evidence that significant differences existed among three scoring results in respect of RMSE (F=25.324, df=2, p<0.001). Post-hoc multiple comparisons suggested significant differences between each pair of scoring decisions ($p_{(multi.MST,Uni.EAP)} = 0.016, p_{(multi.MST,Uni.MIRT)} < 0.001, p_{(Uni.EAP,Uni.MIRT)} < 0.001$). It is worth noting that although in average unidimensional ca-MST with MIRT scoring yielded the smallest RMSE while that with IRT scoring yielded the largest, this did not hold true when attribute correlation varied. When the four measured traits were barely correlated, the scores obtained by multidimensional ca-MST had the largest RMSE, and that obtained by unidimensional ca-MST with MIRT scoring had the smallest.

Table 13 shows the mean residual correlation of scoring results, which was averaged across the lower triangle of the attribute correlation matrix. The residual correlation between two attributes was calculated as the Pearson product-moment correlation coefficient between the estimation residuals of two scales.

$$r_{e_X.e_Y} = \frac{COV(e_X,e_Y)}{\sigma_{e_X}\sigma_{e_Y}}, \tag{17}$$

where $e_x = \hat{\theta}_X - \theta_X$, $e_Y = \hat{\theta}_Y - \theta_Y X \in \{1, 2, \ldots, K\}$.

Again, lower residual correlation was obtained when item pools were informative and moderate difficult. When there were little or moderate correlation among traits, MIRT scoring – under both the multidimensional and the unidimensional ca-MST

scenarios - produced smaller residual correlation than IRT scoring. However, when measured traits were highly correlated, employing unidimensional ca-MST with separate IRT scoring consistently yielded the smallest residual correlation.

These results suggest that the multidimensional ca-MST mode may not be a better alternative to its unidimensional version under certain scenarios, due to its high RMSE when traits do not correlate, and high residual correlation when traits correlate high. When the correlations among traits are so high that multiple traits almost point to the same latent scale, MIRT is unfavorable to model the latent space, and unsurprisingly, does not efficiently extract the information, which leads to a high residual correlation matrix.

Table 12. Mean RMSE of Scoring Results with Multidimensional and Unidimensional ca-MST

r	Configuration	a	b	Multi.MST	Uni.IRT	Uni.MIRT
0.2	1-3	1	-1	.529	.512	.507
			0	.511	.477	.473
		.6	-1	.644	.631	.622
			0	.620	.612	.605
	1-2-3	1	-1	.459	.438	.435
			0	.426	.401	.399
		.6	-1	.562	.554	.548
			0	.542	.533	.528
	1-3-3	1	-1	.459	.427	.424
			0	.443	.394	.391
		.6	-1	.565	.547	.541
			0	.549	.527	.522
0.5	1-3	1	-1	.503	.512	.483
			0	.485	.478	.452
		.6	-1	.603	.628	.585
			0	.584	.612	.571
	1-2-3	1	-1	.437	.437	.416
			0	.409	.402	.384
		.6	-1	.532	.555	.520
			0	.514	.534	.502
	1-3-3	1	-1	.438	.426	.406
			0	.423	.394	.378
		.6	-1	.534	.548	.514
			0	.518	.527	.497
0.8	1-3	1	-1	.438	.512	.423
			0	.422	.477	.397
		.6	-1	.521	.631	.508
			0	.502	.612	.494
	1-2-3	1	-1	.385	.437	.368
			0	.361	.401	.344
		.6	-1	.460	.555	.451
			0	.444	.533	.437
	1-3-3	1	-1	.388	.426	.362
			0	.371	.393	.339
		.6	-1	.461	.547	.445
			0	.447	.527	.432
Mean				.486	.504	.464
Sd				.071	.076	.075

Table 13. Mean Residual Correlation of Scoring Results with Multidimensional and Unidimensional ca-MST

r	configuration	a	b	Multi.MST	Uni.IRT	Uni.MIRT
0.2	1-3	1	-1	.049	.058	.046
		1	0	.047	.052	.041
		.6	-1	.075	.084	.071
		.6	0	.069	.078	.065
	1-2-3	1	-1	.039	.047	.035
		1	0	.033	.042	.029
		.6	-1	.056	.067	.054
		.6	0	.052	.063	.050
	1-3-3	1	-1	.037	.043	.030
		1	0	.035	.041	.029
		.6	-1	.056	.064	.051
		.6	0	.055	.062	.050
0.5	1-3	1	-1	.130	.151	.119
		1	0	.121	.132	.103
		.6	-1	.190	.208	.178
		.6	0	.175	.197	.166
	1-2-3	1	-1	.097	.117	.086
		1	0	.085	.105	.074
		.6	-1	.145	.169	.138
		.6	0	.133	.156	.125
	1-3-3	1	-1	.098	.114	.082
		1	0	.088	.103	.073
		.6	-1	.147	.166	.135
		.6	0	.134	.153	.122
0.8	1-3	1	-1	.320	.258	.299
		1	0	.294	.220	.258
		.6	-1	.420	.340	.402
		.6	0	.391	.318	.379
	1-2-3	1	-1	.255	.205	.227
		1	0	.217	.182	.196
		.6	-1	.339	.279	.328
		.6	0	.312	.252	.304
	1-3-3	1	-1	.258	.199	.218
		1	0	.229	.170	.188
		.6	-1	.339	.267	.316
		.6	0	.317	.244	.295

Effect of Attribute Correlation

The effect of attribute correlation on the accuracy and efficiency of a multidimensional ca-MST panel was examined under the assumption that item banks were optimally designed for the specific multidimensional ca-MST configuration. Mean bias, RMSE, and correlations to true thetas were calculated with varied attribute correlation. A 96-item fixed form was simulated as the baseline condition.

Table 14 displays the mean bias of multidimensional ca-MST scoring results with optimal item pools, where μ_a =1, μ_b =0. The amount of bias under each condition is very small. Although there are variations among different condition, these differences may simply attribute to random error. And the factor of attribute correlation does not seem to affect the scoring bias.

Table 14. Mean Bias of Multidimensional ca-MST with Optimal Item Pools

	Configuration	sub-module size			
r	Design	3	5	8	baseline
0.2	1-3	-.001	.000	-.001	-.005
	1-2-3	.003	-.001	.000	
	1-3-3	.004	.000	-.004	
0.5	1-3	.002	.001	-.001	-.002
	1-2-3	.002	-.001	.000	
	1-3-3	.003	.001	-.005	
0.8	1-3	.002	.001	-.002	.003
	1-2-3	.003	-.002	.000	
	1-3-3	.004	-.002	-.004	

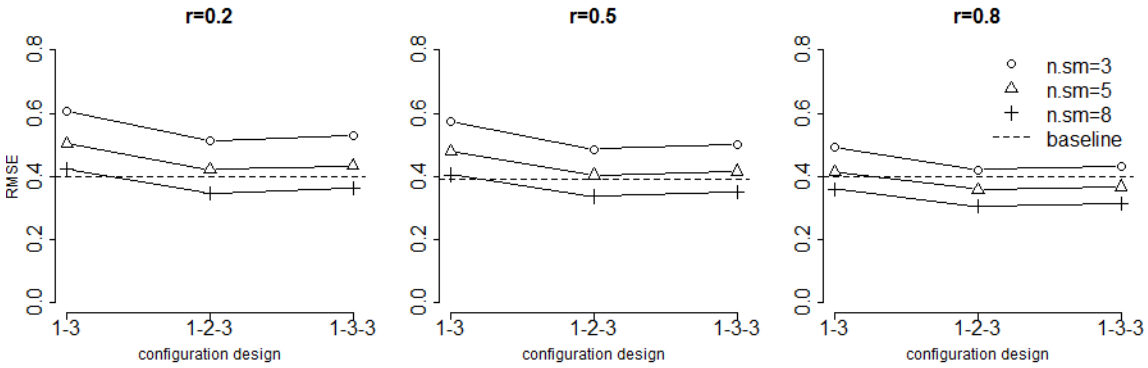
Figure 14 displays a side by side comparison of the RMSE with three attribute correlation levels. The x-axis represents configuration designs, and the y-axis represents the magnitude of RMSE. Three different symbols represent varied lengths of a sub-

module. The dotted lines mark the baseline RMSEs, which do not differ much in the three panels. Figure 14 suggests that the decrease of RMSE corresponds to the increase of attribute correlation. As attributes become more similar, more information carried by one skill can be used to estimate the proficiency of other skills, and thus results in better scoring accuracy. However, the difference between r=0.2 and r=0.5 conditions are not very significant, while r=0.8 decreases the RMSE in considerable amounts.

In addition, the sub-module size plays an important role in determining the RMSE of estimated scores. A larger sub-module size, which corresponds to a longer test, is always associated with smaller RMSE, given the same configuration design. Test length is also determined by the number of adaptive stages. Therefore, the 2-stage 1-3 design consistently yields larger RMSE than its 3-stage counterparts. The performance of the 1-2-3 and the 1-3-3 designs are very similar, although the former yields slightly lower RMSE.

In a 3-stage multidimensional ca-MST design, the longest test – 96 items – is obtained when the sub-module size is 8. Multiplied by four attribute, an examinee takes a 32-item module in each stage. Comparing to the baseline conditions, the multistage tests always achieve better accuracy (lower RMSE) with the same test length.

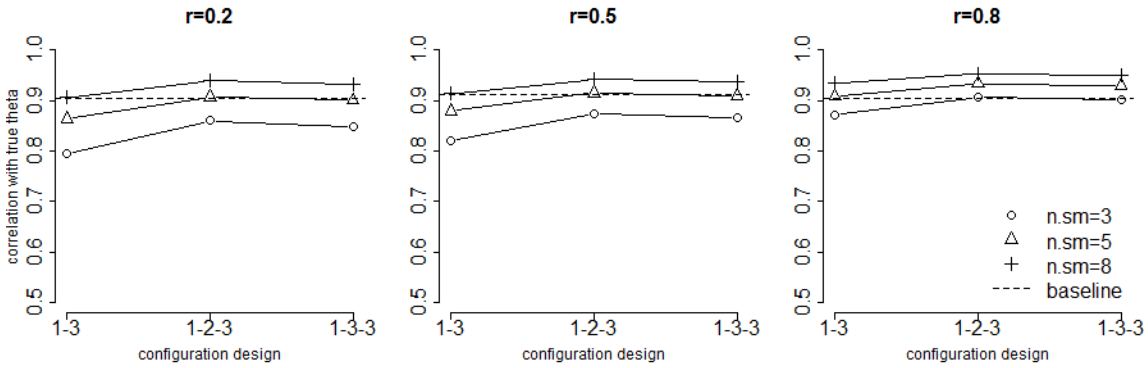
Figure 14. RMSE of Multidimensional ca-MST Scoring with Optimal Item Pool



Similarly, the correlation between the estimated and the true thetas $r_{\theta\theta'}$ under three attribute correlation conditions are presented in Figure 15. As another criteria of estimation accuracy, $r_{\theta\theta'}$ coincides with the RMSE results. Higher attribute correlation is associated with higher $r_{\theta\theta'}$, holding other factors the same. The conditions of r=0.2 and r=0.5 have very similar results, while r=0.8 yields higher $r_{\theta\theta'}$ than the former two conditions.

Given the same configuration design, a larger sub-module size is associated with a higher$r_{\theta\theta'}$. Also, the 3-stage designs yields higher $r_{\theta\theta'}$ than the 2-stage 1-3 design. Again, the performance of the 1-2-3 and the 1-3-3 designs are very similar, although the former yields slightly higher $r_{\theta\theta'}$. Comparing to the baseline conditions, the multistage tests always achieve better accuracy (higher $r_{\theta\theta'}$) with the same test length.

Figure 15. $r_{\theta\theta'}$ of Multidimensional ca-MST Scoring with Optimal Item Pool



The statistics of RMSE and $r_{\theta\theta'}$ that describes estimation accuracy agree with each other. Higher attribute correlation indicates more information shared by the sub-tests, and therefore results in more accurate scoring decision. Comparing to the baseline condition, using multidimensional ca-MST panels can achieve similar or better scoring accuracy with fewer items. When the attribute correlation is low to moderate, the 1-3 design with 8 items in each sub-module, or the 1-2-3 and 1-3-3 designs with 5 items in each sub-module provide similar accuracy with the baseline condition, using only 64 or 60 items. When the attribute correlation goes high, employing the 1-3 design with 5 items in a sub-module, or the 1-2-3 and 1-3-3 deigns with 3 items in a sub-module can achieve similar accuracy. That reduces the test length from 96 items to 40 or 36. Choosing the 1-3 design with 8 items in each sub-module, or the 1-2-3 and 1-3-3 designs with 5 items each sub-module can achieve better accuracy with only 64 or 60 items.

Effect of Item Pool Characteristics

To investigate whether capitalizing on the information carried by the correlation matrix can compensate for a non-optimal item pool, the accuracy and efficiency of multidimensional ca-MST panels were compared between the optimal item pool condition and the three suboptimal item pool conditions. A suboptimal item pool may be low in item difficulty, or item discrimination, or both.

Table 15 displays the mean bias with suboptimal item pools. Compared to Table 14 where an optimal item pool is employed, conditions listed in Table 15 yield higher estimation bias. Testing the main effect of item pool suggests strong evidence that there is difference in mean bias among four item pool conditions ($p<0.001$). Multiple comparisons between each pair of conditions suggest that the optimal item pool condition has smaller bias than any of the suboptimal conditions ($p<0.001$). When the pool is not informative, a moderate pool yields smaller bias than an easy pool ($p<0.001$). However, there is insufficient evidence of difference between an informative pool and an uninformative pool when $\mu_b = -1$ ($p=0.062$). Nevertheless, the magnitudes of bias under the suboptimal item pool conditions are still small, and would not contribute to considerable error in score reporting.

Table 15. Mean Bias of Multidimensional ca-MST with Suboptimal Item Pools

r	Configuration Design	Sub-module Size 3	5	8
	μ_a =1, μ_b = -1			
0.2	1-3	.008	.011	.011
	1-2-3	.009	.008	.012
	1-3-3	.007	.014	.010
0.5	1-3	.006	.009	.012
	1-2-3	.009	.010	.012
	1-3-3	.009	.013	.010
0.8	1-3	.007	.015	.014
	1-2-3	.009	.009	.014
	1-3-3	.007	.012	.011
	μ_a =.6, μ_b = 0			
0.2	1-3	.009	.005	.001
	1-2-3	.005	.004	.004
	1-3-3	.002	.004	.000
0.5	1-3	.010	.009	.005
	1-2-3	.008	.004	.004
	1-3-3	.001	.004	.002
0.8	1-3	.009	.006	.003
	1-2-3	.005	.004	.006
	1-3-3	.001	.004	.000
	μ_a =.6, μ_b = -1			
0.2	1-3	.004	.003	.008
	1-2-3	.010	.011	.010
	1-3-3	.006	.005	.010
0.5	1-3	.009	.003	.010
	1-2-3	.010	.011	.009
	1-3-3	.007	.008	.010
0.8	1-3	.005	.005	.013
	1-2-3	.015	.015	.011
	1-3-3	.011	.011	.010

The mean RMSE of three suboptimal item pool conditions are displayed in Figure 16. Compared to Figure 14 where the item pools are optimally designed, the magnitudes

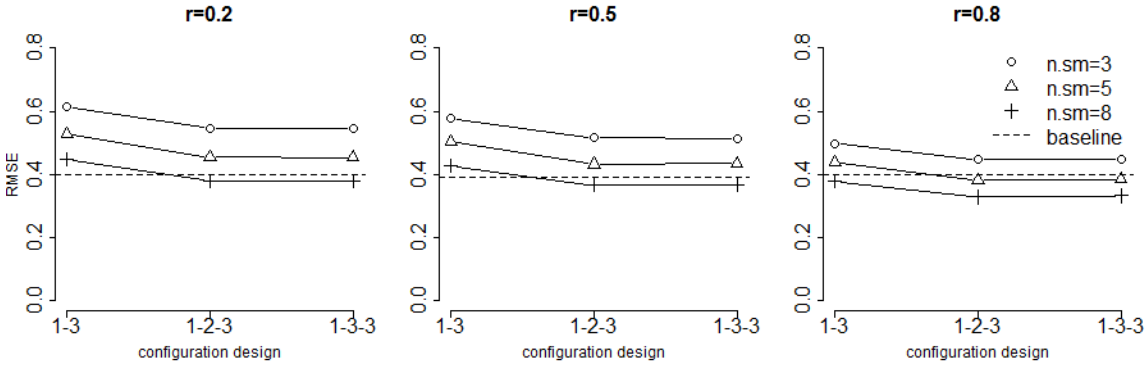
of RMSE in Figure 16 are considerable higher. Both item discrimination and item difficulty affect RMSE. Lower RMSE can be observed where item pools are more informative, or/and with moderate difficulty.

With suboptimal item pools, the efficiency of a multidimensional ca-MST could be substantially affected. With an informative but very easy item pool, a multidimensional ca-MST design can achieve slightly lower RMSE than the baseline condition with the same test length. However, as the item pool becomes uninformative, with the same test length, a multidimensional ca-MST design does not have any benefit over baseline conditions except when the four attributes are highly correlated.

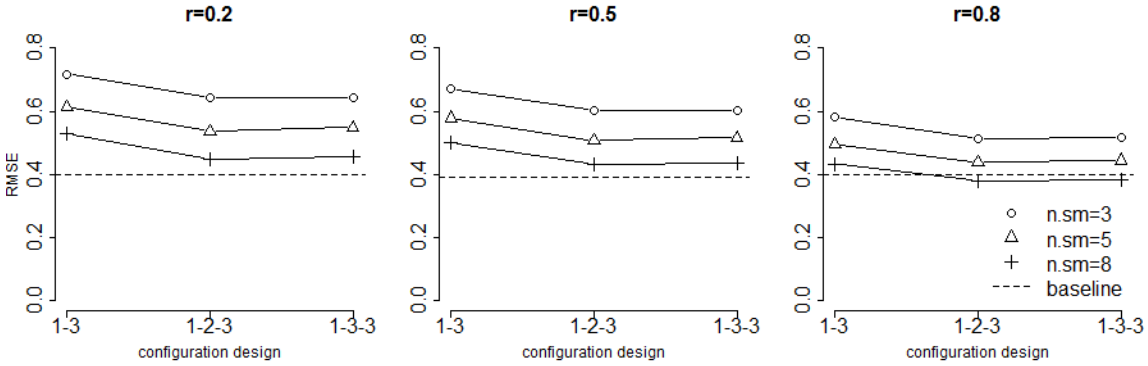
Figure 17 contains the mean $r_{\theta\theta'}$ of three suboptimal item pool conditions. Compared to Figure 15, consistently lower $r_{\theta\theta'}$ are found in Figure 17. It is interesting to note that with certain suboptimal item pools ($\bar{a}$ = 1, $\bar{b}$ = -1), a $r_{\theta\theta'}$ higher than the baseline can still be obtained with the same test length. When r=0.8, similar or higher $r_{\theta\theta'}$ could be obtained with even fewer items (e.g. 1-2-3 design with 5 items in each sub-module, that is, a test length of 60). However, when the item pool is not informative, a higher-then-baseline $r_{\theta\theta'}$ can only be achieved when r=0.8 with a 96 item multidimensional ca-MST panel.

Figure 16. Mean RMSE of Multidimensional ca-MST with Suboptimal Item Pools

μ_a =1, μ_b = -1



μ_a =.6, μ_b = 0



μ_a =.6, μ_b = -1

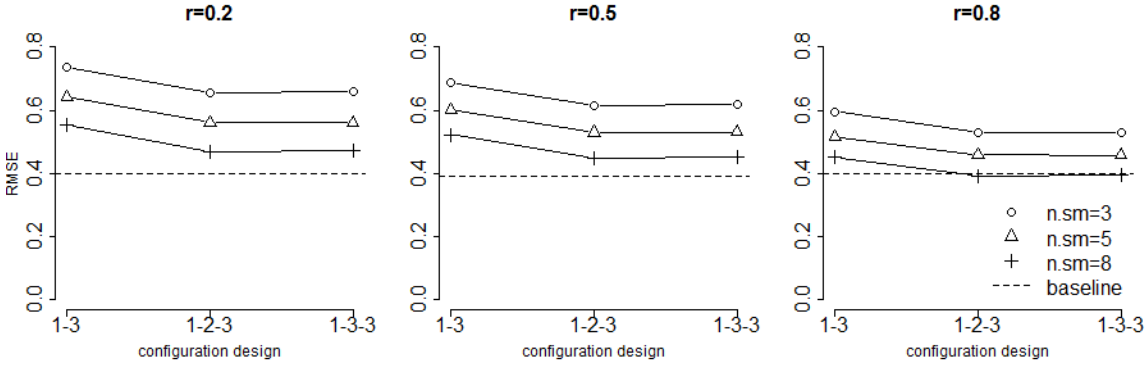
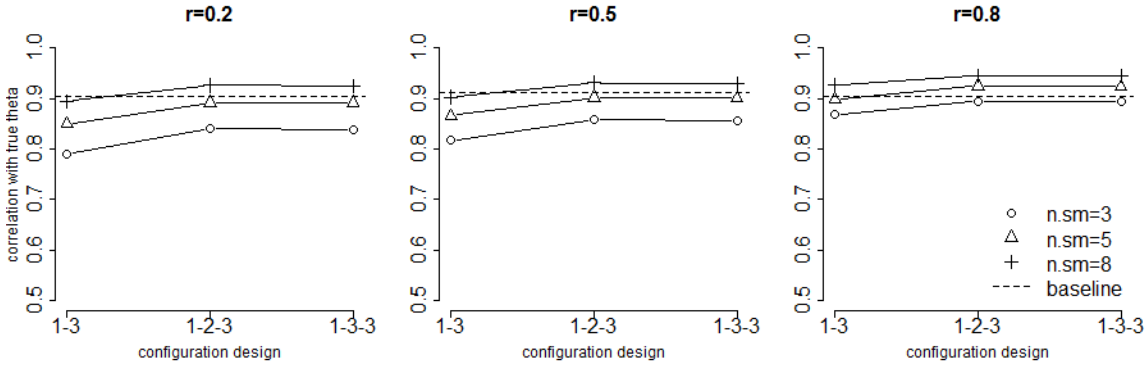
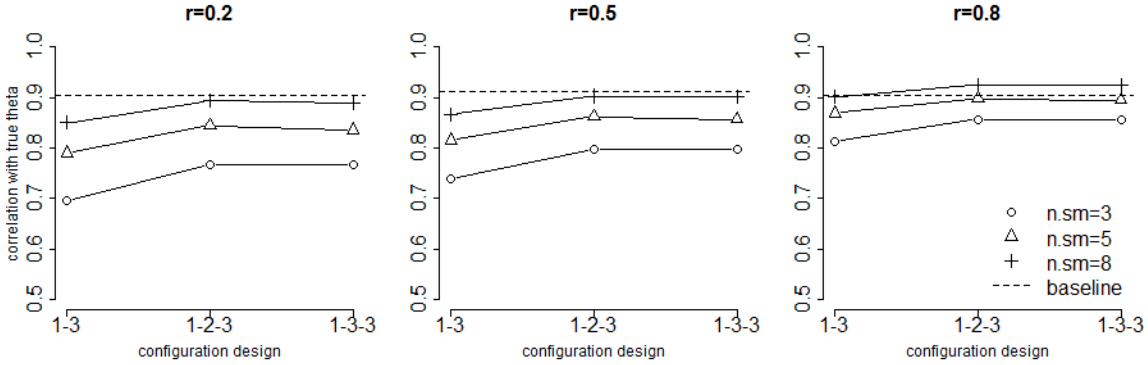


Figure 17. $r_{\theta\theta'}$ of Multidimensional ca-MST Scoring with Suboptimal Item Pool

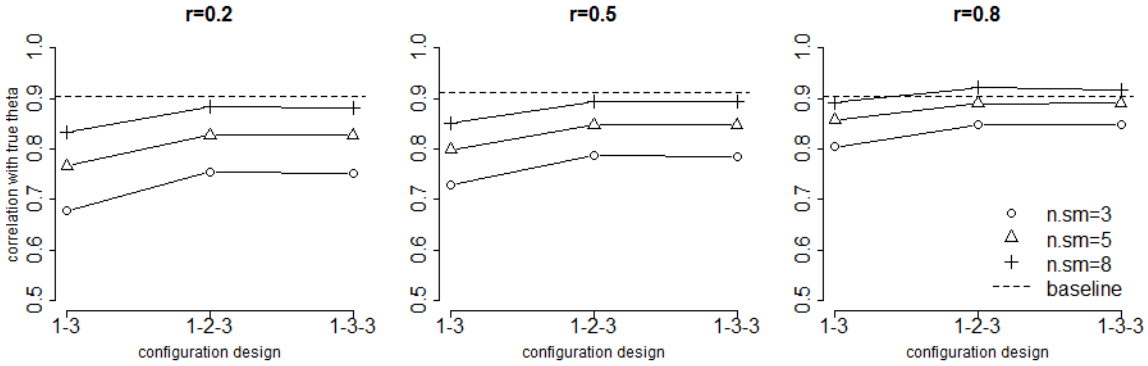
μ_a =1, μ_b = -1



μ_a =.6, μ_b = 0



μ_a =.6, μ_b = -1



Summarizing the above results of mean bias, RMSE and $r_{\theta\theta'}$, item pool characteristics largely determines the accuracy and efficiency of multidimensional ca-MST panels. The benefit of employing a multidimensional ca-MST can be maximized only when the item pool is optimally constructed for the specific configuration design. An inappropriate location of pool difficulty or insufficient item pool information renders the estimation more biased, and less accurate (reflected by higher RMSE and lower correlation with true thetas). It seems that the information of an item bank is more critical than the difficulty location. Suggested by the first row of Figure 16 and 17 where the item bank is informative but too easy, a multidimensional ca-MST panel could compensate for the limit of item bank to some extent, especially when the multiple measured skills are highly correlated. On the other hand, when an item bank is not informative, the corresponding accuracy of a multidimensional ca-MST design may be even worse than a non-adaptive test.

Find an Optimal ca-MST Design

Table 14 and 15, along with Figure 15 to 17 indicate that the effect of attribute correlation is consistent across four item pool characteristic conditions. That said, an optimal multidimensional ca-MST panel can be achieved only with an informative item pool that locates appropriately at the ability scale.

Comparing among three configuration designs addressed in this study, the 1-2-3 design is the best choice in terms of accuracy and efficiency. This design is also most promising when practical considerations are taken into account. To achieve the

comparable performance of the 1-2-3 design, the two-stage 1-3 design needs a longer test. In addition, the two-stage solution only allows one adaptation, which subject to large error if an examinee did not perform well in the beginning of the test due to any factor (e.g. psychological pressure) other than the measured skills. On the other hand, although the 1-3-3 design performs almost as good as the 1-2-3 design, it requires a larger item bank because more alternative modules are needed in the second stage.

CHAPTER V

CONCLUSIONS

Summary and Implication of the Results

As an alternative method of providing sub-scores, multidimensional ca-MST panels (multi.MST in short) were compared to a sequential of separate unidimensional panels in this study. Using the IRT scoring in a unidimensional ca-MST (uni.IRT) represents the method of current practice. In addition, employing the MIRT model (uni.MIRT) matches the scoring method in multidimensional ca-MST. In doing so, the only difference between the uni.MIRT and multi.MST only lied in the selection of modules.

When the measured attributes merely correlate, a uni.IRT outperforms a multi.MST in terms of estimation accuracy. This is expected because a multi.MST is not able to capitalize on the correlation matrix if there is barely any correlation. On the other hand, as the attribute correlation increases to 0.5 or above[4], a multi.MST becomes more accurate than a uni.IRT. It seems that a uni.MIRT almost always provides better estimation results than multi.MST although it is slightly more biased. A possible reason for the suboptimal performance of multi.MST is that MIRT scoring draws the subscores toward their mean across attributes, which results in the loss of some accuracy at the two

[4] This study only has three correlation levels: 0.2, 0.5 and 0.8. However, a multi.MST may start outperforming a uni.IRT with attribute correlations higher than 0.2 but lower than 0.5.

ends of ability scales. When MIRT is used for the routing purpose, an examinee may not be routed to an optimal module.

Besides statistical evidence, the advantage of a multi.MST should also be evaluated in respect of test administration. If four attributes are measured, employing multidimensional ca-MST panels only needs one test, while four tests are required if unidimensional ca-MSTs were to be applied. Under certain scenarios, administering one test for multiple attributes is more reasonable. For example, we assume an above-average student is taking a math test that measures algebra, geometry, and trigonometry. If the student goes through three three-stage uni.MIRT processes consecutively, s/he is likely to take the M-H-H, M-H-H, M-H-H route. The difficult module in the previous subtest may cost her/him too much time so that the time allowance for later subtests is considerably reduced. Even if the administration time is controlled for each subtest, failing to finish the previous difficult items may cause psychological pressure and affect the later performance on even the medium difficult items. One may argue that these three ca-MST process could be delivered separately. In that case, much more efforts in test administration are needed. On the other hand, implementing a multidimensional ca-MST panel can avoid these problems while scarifying the scoring accuracy slightly. The comparison between a multidimensional ca-MST mode and a sequential unidimensional ca-MST mode, however, does not mean to recommend a "better method". Instead, both test mods are promising alternatives for multidimensional assessment when simple structure assumption is held.

The correlation among attributes determines the information shared by multiple scales. Higher correlation indicates more information that a scale can provide for the estimation of the other(s). Therefore, in a multidimensional ca-MST panel, higher attribute correlation is associated with higher accuracy and better efficiency. It is interesting to notice that the difference between the r=0.2 and the r=0.5 condition is much smaller than that between r=0.5 and r=0.8. This suggests the effect of attribute correlation is not linear. Also, it seems that the benefit of MIRT scoring is more prominent with high attribute correlation. Nevertheless, this research does not thoroughly study all levels of correlation. It is still possible that a medium level of correlation (e.g. 0.6) can distinguish itself well from low correlation conditions, although not shown in this study.

Item pool characteristic is another factor that considerably impacts the accuracy and efficiency of a multidimensional ca-MST panel. As mentioned earlier, an informative and appropriately located item pool can optimize the benefit of a multidimensional ca-MST panel. It seems that the information of an item pool is more critical. When the information is adequate, the multidimensional ca-MST design can compensate for an easy pool in some degree. On the contrary, even with appropriate difficulty, an uninformative item pool cannot support an efficient multidimensional ca-MST design. This indicates that when constructing an item pool, items with low a parameters may be abandoned.

The results in this study suggest the 1-2-3 configuration as the best choice among three studied designs in respect of accuracy, efficiency and practical considerations. It

should be noted that this conclusion is obtained with the specific test construct methods applied in this study (e.g. AMI for finding TIFs, backward assembly of sub-modules). It is possible that other test assembly strategy may construct sub-modules with different information curve patterns. In that case, the results may or may not hold true.

Limitations and Future Research

The present study is the first step of exploring the accuracy and efficiency of multidimensional ca-MST panels. To keep the study in a manageable scope, factors and variable levels considered in this study is limited. First, the level of sub-module size meant to represent a reasonable range of test length. However, the choice of levels was not thorough and did not rely on previous researches and was somewhat arbitrary. It is the same problem with the factor of attribute correlation. Although the magnitudes of correlation coefficients represent low, medium and high correlation conditions, the choice of number, again, was arbitrary and incomplete. Future studies can investigate in all levels of these two factors. However, to incorporate complete levels of all variables at the same may result in a very large and time-consuming simulation. Therefore, a better strategy would be to conduct thorough research on each factor respectively and find out the most meaningful levels.

Second, we were interested in how the MIRT scoring in a multidimensional ca-MST can compensate for suboptimal item pools. The current study only compared two levels of the discrimination parameter. The four sub-pools were assumed to be of equal quality. It would be more insightful if some sub-pools were informative while the others

were not. In that case, we may be able to see how the four attributes compensate each other.

Moreover, the statistics (bias, RMSE, etc.) of all test modes here was calculated based on the entire sample. Previous study has suggested uneven distribution of measurement error across an ability scale. It is very common that examinees at the two ends of an ability scale are not estimated as well as those in the center. Wang, Fluegge, and Luecht (2012) demonstrated that when the item pool was optimally designed for a ca-MST panel, the measurement error was consistently small across an ability scale. However, this may not hold true in the multidimensional cases of the current study, especially in the scenarios where suboptimal pools were employed. Future research should examine the measurement error by ability groups so that we can make sure that the design we proposed well serve all test takers.

When the measured attributes were highly correlated, the residual correlation was obviously non-zero no matter which test mode was employed. The logic next step would be to compare the multidimensional assessments discussed in the current study with a unidimensional assessment that covers four attributes. It is possible that when attributes correlates too high they point to the same scale, and a unidimensional assessment may be more appropriate.

REFERENCES

Birnbaum, A. (1968). Test scores, sufficient statistics, and the information structures of tests. In Luecht, R. M. and Nungester, R. J. (1998). Some Practical Examples of Computer-Adaptive Sequential Testing. *Journal of Educational Measurement*, 35(3), 229-249.

Bloxom, B.,&Vale, C.D. (1987).Multidimensional adaptive testing: An approximate procedure for updating. In Meeting of the psychometric society.Montreal, Canada, June.

Brown, J., & Weiss, D. (1977). *An Adaptive Testing Strategy for Achievement Test Batteries*. Contract No. N00014-76-C-0627, NR150-389. Office of Naval Research

Chalhoub–Deville, M., & Deville, C. (1999). Computer Adaptive Testing in Second Language Contexts. *Annual Review of Applied Linguistics*, *19*, 273–299.

Cheng, Y. (2009). When Cognitive Diagnosis Meets Computerized Adaptive Testing: CD-CAT. *Psychometrika*, *74*(4), 619–632.

Dallas, A., Wang, X., Furter, R., & Luecht, R. (2012). Item Pool Size , Targeted Item Writing , and Panel Replication Strategies for a 1-3-3 Multistage Test Design. In *Annual Meeting of the National Council on Measurement in Education* (pp. 1–23). Vancouver, British Columbia, Canada.

de la Torre, J., Song, H., & Hong, Y. (2011). A Comparison of Four Methods of IRT Subscoring. *Applied Psychological Measurement*, *35*(4), 296–316.

de la Torre, J. (2008). Multidimensional scoring of abilities: The ordered polytomous response case. Applied Psychological Measurement, 355-370.

de la Torre, J., & Patz, R. J. (2005). Making the most of what we have: A practical application of MCMC in test scoring. Journal of Educational and Behavioral Statistics, 295-311.

Haberman, S. J. (2008). When can subscores have value?. *Journal of Educational and Behavioral Statistics*, 33(2), 204-229.

Hambleton, R. K., & Xing, D. (2006). Optimal and Nonoptimal Computer-Based Test Designs for Making Pass – Fail Decisions. *Applied Measurement in Education*, *19*(3), 221–239.

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.

Lord, Frederic M. *Applications of item response to theory to practical testing problems*. Lawrence Erlbaum, 1980.

Lord, F. (1968). Some Test Theory for Tailored Testing. Princeton, NJ: Educational Testing Service.

Luecht, R. (2012). A Technical Design Note on Computer-Adaptive Multistage Testing (ca-MST) Panels for Multidimensional Assessments. An iternal technical report.

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

Luecht, R., Brumfield, T., & Breithaupt, K. (2006). A Testlet Assembly Design for Adaptive Multistage Tests. *Applied Measurement in Education*, (19 (3)), 189–202.

Luecht, R. (2003). Multistage Complexity in Language Proficiency Assessment : A Framework for Aligning Theoretical Perspectives , Test Development , and Psychometrics. *Foreign Language Annals*, *36*(4), 527–535.

Luecht, R. & Burgin, W. (2003). Test Information Targeting Strategies for Adaptive Multistage Testing Designs. In *Annual Meeting of National Council on Measurement in Education*. Chicago, IL.

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

Luecht, R. (1998). Computer-Assisted Test Assembly Using Optimization Heuristics. *Applied Psychological Measurement*, *22*(3), 224–236.

Luecht, R. (1996). Multidimensional Computerized Adaptive Testing in a Certification or Licensure Context. *Applied Psychological Measurement*, *20*(4), 389–404.

Luecht, R. M., & Hirsch, T. M. (1992). Item selection using an average growth approximation of target information functions. *Applied Psychological Measurement*, 16(1), 41-51.

Mulder, J., & van der Linden, W. (2009). Multidimensional Adaptive Testing with Optimal Design Criteria for Item Selection. *Psychometrika*, *74*(2), 273–296.

Reckase, M. (1974). An interactive computer program for tailored testing based on the one parameter logistic model. *Behavior Research Methods and Instrumentation* 6 (2), 208-212.

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

Sinharay, S., Puhan, G. and Haberman, S. J. (2011), An NCME Instructional Module on Subscores. *Educational Measurement: Issues and Practice*, 30: 29–40

Sinharay, S. (2010). How often do subscores have added value? Results from operational and simulated data. *Journal of Educational Measurement*, 47(2), 150-174.

Stone, C., Ye, F., Zhu, X. & Lane, S. (2009). Providing Subscale Scores for Diagnostic Information: A Case Study When the Test is Essentially Unidimensional. *Applied Measurement in Education*. 23(1), 63-86

van der Linden, W. J., & Luecht, R. M. (1995). An optimization model for test assembly to match observed score distributions. In G. Englehard & M. Wilson (Eds.), *Objective Measurement: Theory into Practice* (Vol. 3). Norwood, N J: Ablex.

Wang, C., Chang, H.-H., & Douglas, J. (2011). Combining CAT with cognitive diagnosis: A weighted item selection approach. *Behavior research methods*, 95–109.

Wang, W. C., Chen, P. H., & Cheng, Y. Y. (2004). Improving measurement precision of test batteries using multidimensional item response models. Psychological Methods, 116-136.

Wang, X., Fluegge, L., & Luecht, R. (2012). A Large-scale Comparative Study of the Accuracy and Efficiency of ca-MST Panel Design Configurations. In *Annual Meeting of the National Council on Measurement in Education*. Vancouver, British Columbia, Canada.

Xing, D., & Hambleton, R. (2004). Impact of Test Design, Item Quality, and Item Bank Size on the Psychometric Properties of Computer-Based Credentialing Examinations. *Educational and Psychological Measurement*, *64*(1), 5–21.

Zenisky, A., Hambleton, R., & Luecht, R. (2010). Multistage Testing: Issues, Designs, and Research. In van der Linden, W. & Glas, C. (Eds.), *Elements of Adaptive Testing* (pp. 355–372). New York, NY: Springer New York.

Zheng, Y., Nozawa, Y., Gao, X., & Chang, H. (2012). Multistage Adaptive Testing for a Large-Scale Classification Test: The Designs, Heuristic Assembly, and Comparison with Other Testing Modes. In *Annual Meeting of the National Council on Measurement in Education*. Vancouver, British Columbia, Canada.