Draft Technical Report

Test Assembly and Delivery Models for a National Examination of Medical Knowledge: Optimal Form Quality and Item Usage from a Modular Design

August 12, 2011

Krista Breithaupt & Donovan Hare

Introduction

The MCQ component of the MCC Qualifying Examination (MCCQE) Part I is administered entirely over the Internet, and uses a fixed length multi-stage computer adaptive design. The test is composed of subject-related groupings of test items by difficulty levels, and branching depends on number-correct scoring during the exam. The design is proprietary to the MCC and has been working well for over a decade.

As adaptive testing became more popular, and ease of communication via technology increased our concern with test security, researchers have focused a great deal of attention on resolving problems related to item over-exposure when item banks are limited and simple heuristics are the basis for item selection. Some operational programs have identified benefits to pre-constructing modules for adaptive or linear administration, while preserving the benefit of increased precision in test scores that can be derived from targeted test construction. The multi-stage or testlet-based model has been adopted by several high-stakes testing programs (e.g. tests for physicians, CPAs, and graduate admissions examinations in the US).

This study is one in a series intended to examine alternative designs for administration of the MCCQE I that could be less dependent on proprietary software, preserve some benefits of the current adaptive design, and take advantage of discrete optimization techniques for automated test assembly. The purpose of the study is to explore practical models for test administration based on a representative item pool from the current exam, using realistic test content and other specifications. This technical report summarizes results from alternative test administration models (linear and modular), and the resulting quality of test forms. Comparability and number of forms, content coverage, and expected score precision are considered as important quality criteria, in addition to item usage and item and form exposures. A discussion is offered on some anticipated benefits of simple linear and multi-stage modular designs for test forms. Some implications for operational implementation are discussed, along with suggestions for future research.

Automated Test Assembly Models

A review of some fundamental concepts and some applications of integer programming and optimization methods from manufacturing are appropriate to the assembly of items into test forms. There are many industrial problems whose solutions require a group of

discrete choices to be made. These choices might take the form of the number of widgets of a given type that should be made and a schedule for the group of machines that make them. The choices usually have natural dependencies that constrain the idealized solution. Perhaps there is an order for some of the machines that build a type of widget or a time delay for a machine to paint widgets with differing colors. These situations are analogous in many respects to our test construction problem.

In the process required to build a test form of traditional multiple-choice questions (MCQs), a viable solution will require us to choose a number of questions from a bank of potential questions. Selection on to test forms is ordinarily guided by, or constrained by, content specifications and other design or business rules, such as form length and item exposure restrictions. In the case that there are many forms to create, and a large bank of items, our objective is to choose items for forms so that the total solution of all tests created is optimal with respect to the most important design goals. That is, the set of forms will be optimal, given the items available and all the design constraints defined in the problem.

When building forms or assembling modules for test delivery, it may be desirable to maximize some function of statistical properties of items to ensure score precision, or to allow for adaptive subtest designs, based on the difficulty of test questions. One example of the importance of statistical properties of items for test and inventory designs is the popularity of the use of item response theory (IRT) for ensuring equivalence across test forms, or in building adaptive subtests (e.g. Luecht & Nungester, 1996). The use of statistical properties of test questions, in addition to the discrete selection variables in the problem, introduces complexity in the overall assembly problem. In the mathematical literature, these kinds of combinatorial problems are modeled as ‘discrete optimization’ problems.

Discrete optimization problems range widely in their difficulty to solve efficiently and in their solution techniques. The structure of the discrete optimization problem solved in assembly of subtests and forms for the MCCQE I makes use of mixed integer programming, an area of study derived from linear programming. Mixed integer programming has been used extensively in a variety of problems from optimal test design to inventory planning (van der Linden, 2006).

Each organization must weigh the potential benefits of alternative administration designs and goals for automated assembly. Some typical considerations in deciding on a preferred approach include the following:

- Equivalence of forms within administrations and across time.
- Flexibility for updating content specifications.
- Efficient and effective use of expert judgment in form creation and approval.
- Minimum item and form exposure within administration periods.
- Even use of the sub-pool or bank of test items.
- Complete coverage of the required knowledge and skills as defined in the test specifications.

- Uniform experiences for test takers (fixed length tests and adequate appointment times).
- High precision in the range of decisions made using the total test score (e.g. decision accuracy at the pass/fail decision point).
- Supports high security of form development and administration schedules (e.g. minimize predictable inventory rotation or real-time item presentations to deter 'gaming' or pre-knowledge of test questions).

The development and design of the multi-stage model currently used for the MCCQE I is described in Blackmore et al. (19XX). More recently, alternatives such as linear on the fly test (LOFT) designs have been explored for the MCCQE I (Wood et al., 2009). To evaluate the potential benefits of some alternative assembly and delivery models for the MCCQE I exam we propose two muti-stage adaptive designs, that may allow efficiencies in test construction while preserving some benefits of the current test administration model.

Methods

A discrete optimization programming language is available in the academic version of the popular CPLEX solver from ILOG (© IBM, 2011). This software is intended to describe and solve linear programming problems where integer and non-integer variables must be represented. In the interactive developer environment, it is possible to describe the objectives of the assembly of items into test modules or forms according to the quality goals of the individual testing organization. The CPLEX solver uses linear programming solutions to examine alternative feasible solutions where different combinations of test items are constructed. Some advantages of this approach to the test assembly problem include the ability to compare feasible solutions, test forms, with the unconstrained 'optimal' problem solution. There are strait forward methods to simplify the multi-factorial variable problem arising from multiple constraints, and to ensure that only feasible solutions are reached that satisfy all of the design rules. The software can run on a typical desktop computer with moderately large memory and processing speed, and produces multiple test forms in only a few minutes.

To create MST forms for the MCCQE I, a constrained problem was described that included rules for the following design elements. There are two steps in the assembly that was conducted for this study, although these can be combined into a single multi-level optimization problem. Two possible adaptive MST models are described as Design A and Design B, and each requires information and decisions about the following basic design constraints:

1. Proportion of test items in each content area (medical discipline).
2. Total test length in items.
3. Number of test items per module.
4. Number of levels of difficulty per stage.
5. Target for total score precision, and for module difficulty.
6. Number of opportunities to tailor to examinee ability (adaptive stages).

7. Restrictions on items that should appear together in a module or form.
8. Available typical sub-pool size.
9. Feasible number of modules, given pool size and content coverage of items.
10. Number of unique forms (or panels composed of modules).

Design A

Since the current MCCQE I is designed as a multi-stage test were there are always a minimum of 144 scored items, a test length of 150 scored items seemed reasonable. Next, the number and length of modules was required, and is also intended to be similar to the current examinee experience. There are seven stages in the current design, and items are grouped into four different difficulty categories. This means there are seven opportunities to adapt to the candidates ability level. The intended benefit of MST is to increase the variety in test forms, target the experience to the level of ability of the examinee, and increase the precision of the total test score and pass or fail decision.

However, there are some potential areas of improvement if we consider alternative designs for the MCCQE I. Because item groupings are defined for each content domain separately (by medical discipline) the actual range of average difficulties for the module is restricted. Also, years of test administrations have shown the population of Canadian trained physicians is very homogeneous. Their average ability level is high, compared to the level of difficulty of the items in the pool. This means that there are only small differences in ability for the majority of test takers, and items measuring along the moderate to higher ability levels in the bank represent only about half of the total available test questions. These factors reduce some benefits of tailoring the examination, since most test takers will see the moderate and high difficulty test items.

To improve on the variety in test forms, Design A is based on modules that include the required content coverage for all disciplines within each module. This should allow the assembly to choose more discriminating items in each module, improving total test score precision. Also, the number of difficulty groupings was reduced to three (easy, moderate and difficult), with targets set to minimize over-use of harder items and boost total score precision near the pass or fail point. A convenient length for modules was set at 35 items each, with no pilot items (these can easily be selected and added in an operational version of the problem).

A sub-pool of 2716 MCQ items from the operational item bank were used for this study. Items were coded to reflect the required medical disciplines, and had more detailed sub-discipline codes based on the published MCC Objectives (cite website). Since the current assembly is done manually, and reviewed iteratively by medical expert panels, there is no coding in the bank currently for item enemies. Our study makes use of the discipline and objective codes to prevent or reduce the recurrence of MCQs on similar topics within any module or panel route.

The structure of the MST is presented in Figure 1. The 140 items are delivered in four blocks of 35 item modules (pre-test items are not included in the study). The first module is of moderate difficulty, and each module exactly represents the required proportion of content for the MCCQE I as described in Table 2 (below). In order to support adaptation, there is a score calculated on the items after each set of 35 items that is used to determine whether a module of equal, greater or lesser difficulty would be selected at the subsequent stage.

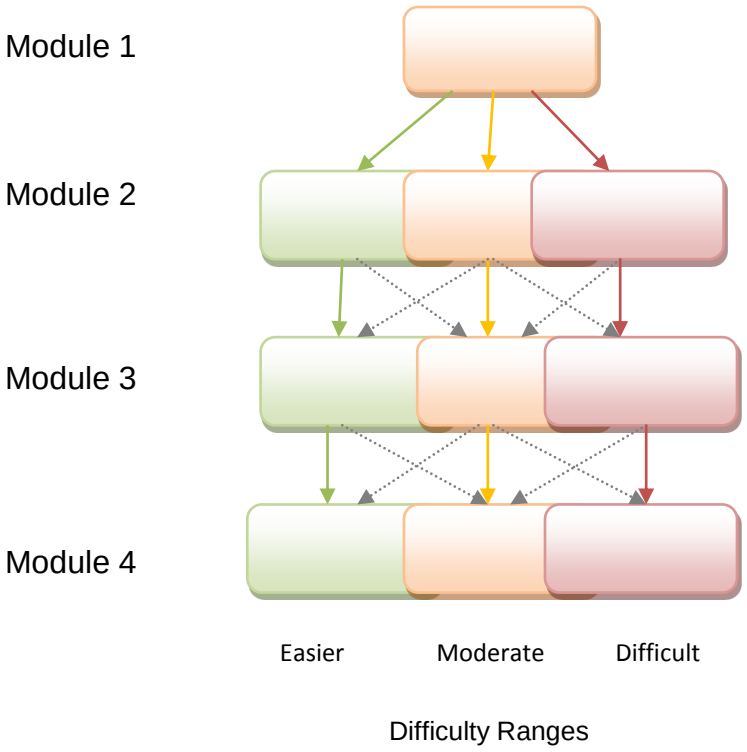

Figure 1. Multi-Stage Test Design A

The modules are each designed to optimize precision at a particular point on the ability scale. In this case, we use the IRT theta scale which is expressed on the same scale as the item difficulty parameters. All items are eligible for selection in the easy, moderate or difficult modules. However, the quality of the solution is constantly re-evaluated as alternative compositions are considered by CPLEX©. This means the result will include the best final solution for maximizing precision at the targeted difficulty level of the module, for all modules.

In order to provide for fairly even use of modules, and high accuracy of decisions made using total scores, the distribution of ability in the population and the location of the pass or fail point were also considered. The population of examinees are of high ability, on average, with a mean on the IRT scale expressed as $\theta = 1.0$. In contrast, the required performance to pass the examination is set at about $\theta = -1.2$ on the same scale using a judgmental and empirically-informed standard setting method. Because the items are

designed to assess basic medical knowledge and skills expected of a Canadian Medical graduate, most items are of a difficulty in the range of the passing score. The bank item difficulty mean is about θ = -1. This means many candidates will be routed to the moderate and difficult modules, although most of the items in the ban, and our most important score accuracy are at the easier to moderate range of test items.

A weighted objective function was defined in the assembly problem to ensure that more discriminating items would be favored in the easier and moderate modules, given all available items. Targets for optimal discrimination (based in module test information functions) were set as describe in Table 1. These are relative weights (judgmental) and both weights and targets differ somewhat for Design B, which is described in the next section.

Table 1: Weights and Targets for Designs A and B

Module	Design A (2718 items)		Design B (1500 items)	
	Target	Weight	Target	Weight
Easier	-1.5	0.8	N/A	N/A
Moderate	-1.0	1.2	-1.5	0.8
Difficult	0.0	0.8	0.0	1.0

Table 2: Content Specifications for MCCQE I

Disciplines	MED	OBGYN	PEDS	PSYCH	SURG	PopH	CLEO	FM
% Weight Overall	16	17	17	17	16	4	13	65*
# Items (Midpoint)	22.4	23.8	23.8	23.8	22.4	5.6	18.2	91
# Per 35 Item Testlet	7.84	8.33	8.33	8.33	7.84	1.96	6.37	22.75
Range +/- 5%	1.12	1.19	1.19	1.19	1.12	0.28	0.91	2.275
Range +/- 10%	2.24	2.38	2.38	2.38	2.24	0.56	1.82	2.275

*Family medicine (FM) is coded across other disciplines, so this is a test-wise weight with non-family medicine items at 35%.

Based on the optimization procedure, where reasonably high discriminating items are selected for modules, the routing for most candidates is expected to follow the primary pathways. Primary pathways can be described as routes where there is no change in module difficulty after the second stage (e.g. moderate to moderate at a given routing stage). A small proportion of candidates would be re-routed to an adjacent module between stages 2 and 4, and an even smaller proportion would re-route before the final stage. This is predictable because the aggregated score during the examination is used at each routing point, and as the test progresses that score becomes a more accurate estimate of the candidate's true ability level.

The MCCQE I is currently calibrated using the 2-parameter Item Response Theory (IRT) model for scoring (the c, or guessing parameter, is fixed at zero). In this study, the IRT parameters are used to select items for modules and to calculate optimality in the discrete programming assembly solution. More detail on the specifics of the general MST model using discrete optimization is provided in Breithaupt & Hare (2007).

The current administration schedule for the MCCQE I allows for testing twice a year over a period of several weeks. Multiple forms are created for each administration period to reduce the exposure of any test question. Approximately 3,000 examinees take the exam in the fall, and about another 1,000 are tested in the spring session.

Design A was constructed to yield 10 unique panels (forms), in order to ensure a wide variety of different modules are seen by examinees during any administration. Forty modules were constructed, with 12 each for easy and difficult and 16 moderate. This was intended to allow sufficient modules per panel. Items were uniquely assigned to modules in the final solution (no re-use of items for any module). A pool of 2718 items were used in this study from the operational bank for the MCCQE I. The panel assembly step was completed after the modules are created, and the optimization is expressed to minimize the re-use of any module across panels and the number of potential item-enemy pairs in any primary route.

Design B

An alternative design recognizes the homogeneity of the majority of our Canadian test-takers. As is noted above, most candidates would be routed to a module of equivalent difficulty as the test progresses. Without specific models for ‘gaming’ or attempting to use the design to improve a total score estimate, we anticipate that there would be a very small amount of re-routing by stage 3, and none by stage 4. This suggests that we might benefit from constructing fewer modules at each stage, and still preserve the benefits of presenting different forms and modules to examinees while maximizing the precision of total scores and the pass or fail decision.

In order to take advantage of the expectation that most test-takers would be routed on the moderate and difficult primary pathways, it seemed reasonable to consider an MST with only two options at each stage. Therefore, Design B consists of a total of 7 modules, also of 35 items for each of the 10 panels (Figure 2). Total test length and the number of panels remains the same for both designs.

Also, it is customary in very high security testing programs that a smaller sub-pool of available items would be considered for any individual administration. In order to represent a condition where fewer items would be used for a given assembly, only 1500 test items, randomly drawn from the available 2718, were used in Design B.

In order to make the best use of the available items in the pool, new targets and weights were set for Design B, and are described in Table 1. The same content specifications were used in this design (Table 2), as were the item enemy definitions.

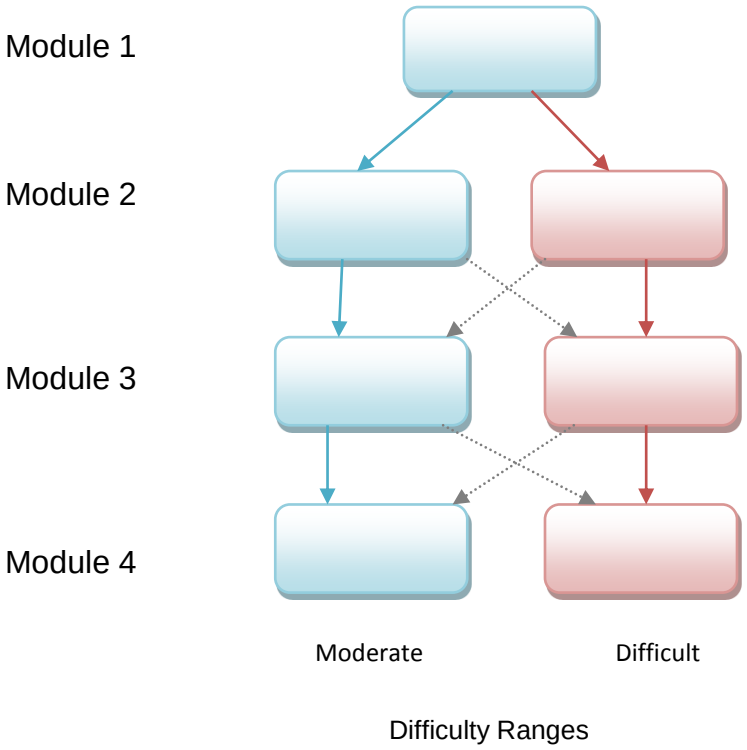

Figure 2. Multi-Stage Test: Design B

Results

Design A: The ILOG application was run on a fairly typical configuration for a desktop system. This was a 32-bit operating system, with a typical 2.8 GHz dual core processer and 4 GB of RAM. The module assembly problem resulted in over 1 million constraints across 100,000 binary and integer variables. A feasible solution was obtained in the module assembly step within 35 seconds, and an optimal solution required a total of 85 seconds to reach optimal (default settings for mathematical programming were used).

The 40 modules were then provided as input to the panel assembly step, and resulted in a mixed integer problem with 128,000 constraints and 19,000 variables. A feasible solution was obtained after 65 seconds. Performance tuning when we coded the constraint model for the panel design identified problems with the strict enemy rule, and this was relaxed for the easy and difficult routes. This means that two items having the same medical discipline and MCC Objective codes could exist across the four panels of most routes. In the absence of a policy and bank review for accurate enemy item coding, this relaxation seemed reasonable. Also, the mixed integer solution tolerance was set at 10% of optimal to deliver a reasonable solution with less processing time that

would normally be used under default settings. The panel assembly and writing output describing each panel required a total of about 1.5 minutes.

The content coverage of medical disciplines was equal and met requirements for all modules, with an acceptable variation of +/- 2 items from the midpoint in Table 1. The maximum reuse of any module across panels was 5 (for moderate), and 3 for easy and difficult. With relaxation of the number of potential pairs of item enemies (based on our strict proxy definition), we obtained a maximum of 63 pairs for any route. This is a minimized function in the assembly; the total possible number of enemy pairs is over 10,000 for this number of items and modules.

A useful convention in high-stakes examinations uses a maximum exposure rule of 10% of test takers as the maximum who could see the same test content. In this build, only 10% would see the same panel, and a smaller proportion would see the same modules. If the ability in the population places most individuals on the moderate or high difficulty permanent routes, the maximum exposure for any module or item would fall to about 5%.

Modules were examined using plots of the Test Information Functions, and appeared very uniform and conformed to the desired targets (Figure 3). The impact of the relative weights is evident when we note the maximum information for modules that are centered on the left (difficult) is lower, compared to those centered on the right (easier). This is a consequence of the greater number of items in the bank at the lower difficulty level (nearer the passing score), and means the selection had more highly discriminating items to choose from in the bank. It is also a sign the assembly meets the intended goal of maximizing information around the passing score, when the moderate modules are peaked in the area around $\theta = -1.0$.

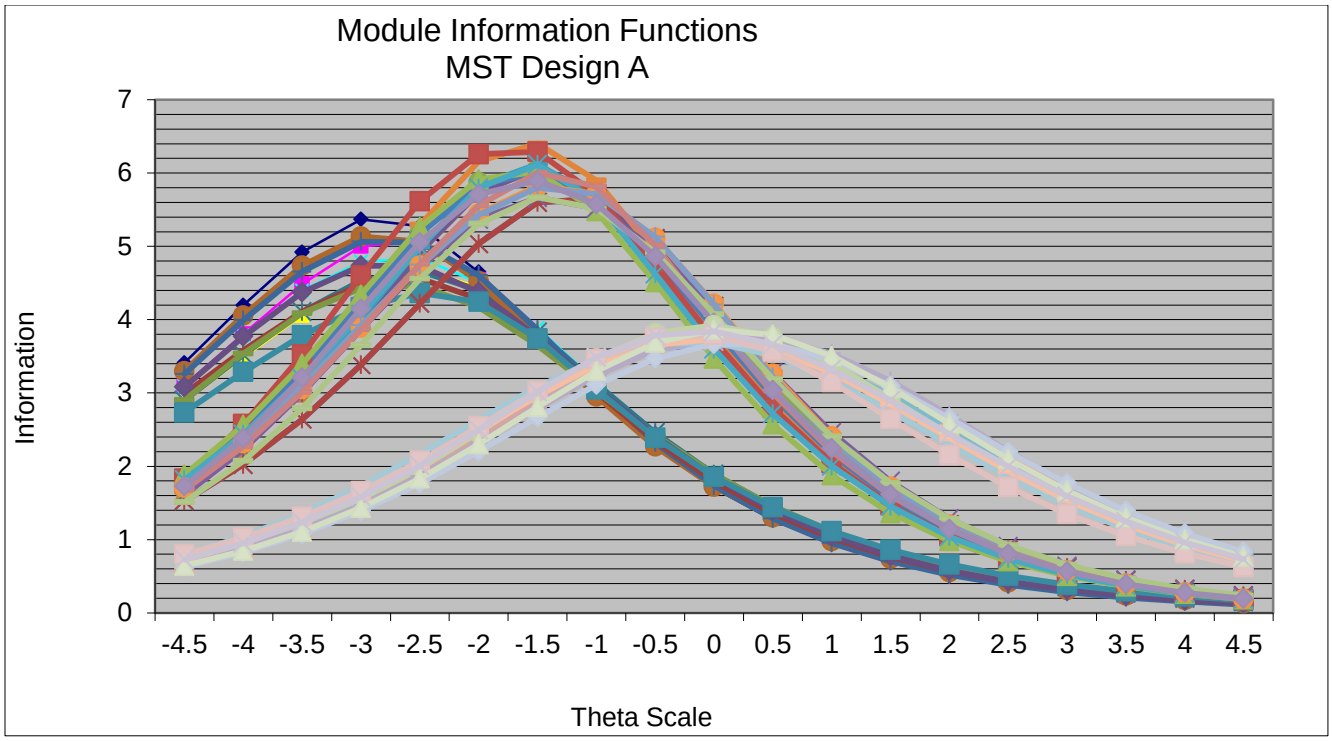

Figure 3: Test Information Functions, Design A

The precision of total scores depends on the discrimination of items at an important point on the ability scale. IRT information is the 'square of the ratio of the slope of the regression of y on theta to the standard error of measurement for a given theta' (Hambleton & Swaminathan, 1985, p.103). As an example of the total test information from this assembly solution, we can look at one example panel in from the results of Design A. All ten panels are very similar, and the total information at the passing score from Panel 3 is depicted in Table 3.

Good precision at the pass or fail score should show up as high and flat curves over the theta point of -1.0 in Figure 3. Since any examinee sees four modules, it is necessary to sum the IRT information for all modules on the path taken. Total score information values at the passing score range from about 15 (easy or difficult route) to 22 (moderate route). Since the moderate route includes the required passing score, we can see the precision is optimal for the pass or fail decision.

Relative interpretations are also made using IRT information. An efficiency gain can be calculated for the impact of the moderate route compared to the other primary routes, at the passing score. The formula is a simple ratio of their information values (Hambleton & Swaminathan, 1985, p. 121):

Relative efficiency of easy route compared with moderate route is calculated as 15/22 = 0.68. This can be interpreted to mean that the precision of scores at the passing point derived from the moderate route would only be obtained if the length of the test for those taking the easy route were increased by 68% (95 items, in addition to the 140). Or, that the optimization of modules, and the routing introduces a significant gain in precision, at least for those taking the moderate route compared with those taking the easier route.

Table 3: Total Score IRT Information Design A

Panel 3	Stage	Easy Route	Moderate Route	Difficult Route
	1*	Module 15(M)	Module 15(M)	Module 15(M)
	2	Module 3(E)	Module 19(M)	Module 31(D)
	3	Module 5(E)	Module 21(M)	Module 33(D)
	4	Module 9(E)	Module 22(M)	Module 37(D)
Precision at Passing (-1.0)		14.99	22.35	15.57

*Note the stage 1 module is the 'routing' module and is the same for each candidate assigned to this panel. A total of 9 different modules are used for one panel.

Using Panel 3 as an example, we can also look at how often enemy items may occur Recall, the primary route for the moderate modules would mean modules 15, 19, 21 and 22 are delivered. In this solution this route includes 43 potential item enemy pairs based on the strict proxy of same medical discipline and MCC Objective.

Design B.

The simplified panel design proved to be a somewhat easier linear programming problem with 161,000 constraints and 32,000 variables. Recall, a smaller sub-pool of 1500 items was used in this design option. The same total test length of 140 items was modeled by creating 35 item modules for each of the four stages of the MST. There were 14 moderate and 10 difficult modules simultaneously assembled for panel placement (24 modules in total). A feasible solution was identified in 63 seconds, and an optimal solution was found in less than 1.5 minutes. The tolerance gap for optimization was also set at 10% for this assembly run.

The same rules for content coverage and enemy-item coding were specified, and were satisfied, as with Design A. The resulting modules were peaked around their intended targets, and very comparable as depicted by the test information functions in Figure 4.

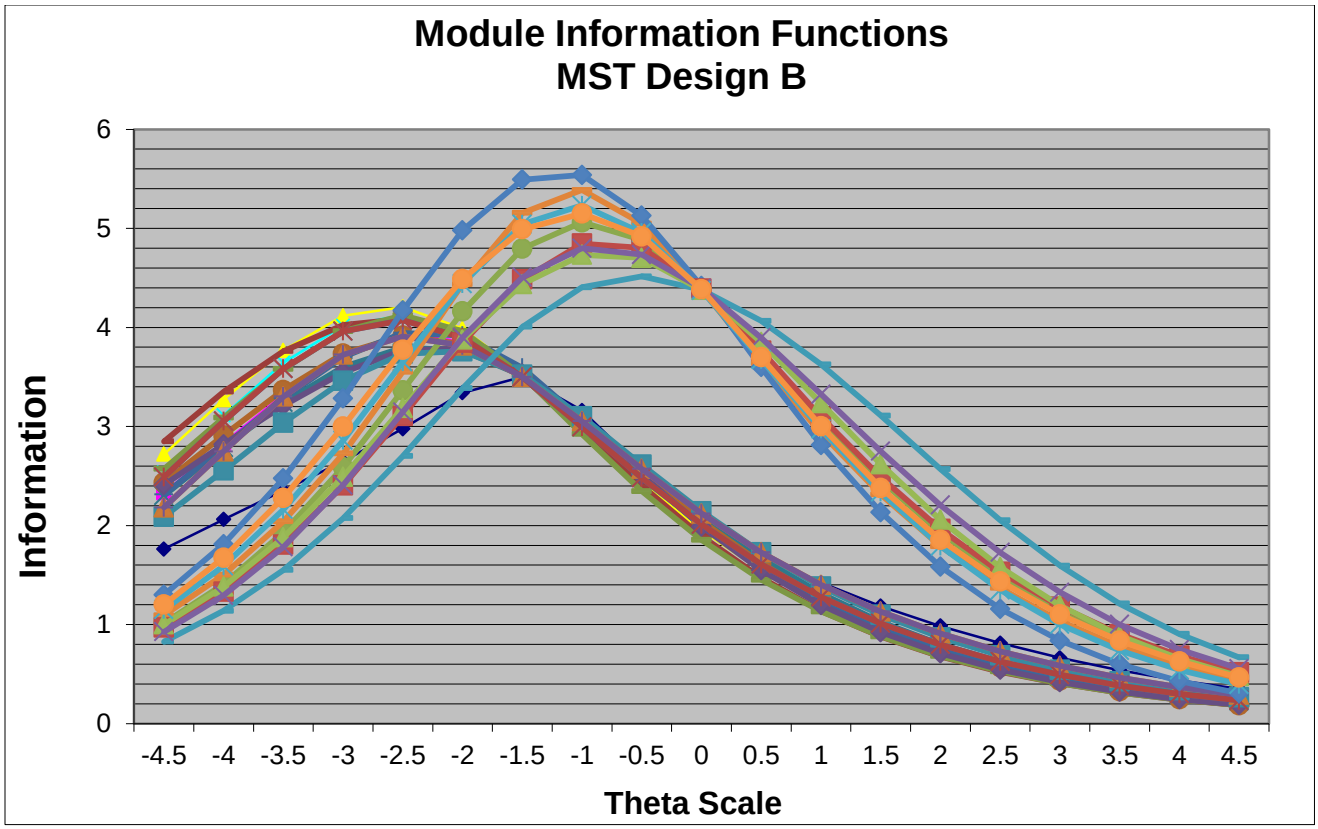

Figure 4: Test Information Functions Design B

Talk about score and test precision for Design B here……and relative efficiency of routes.

Table 4: Total Score IRT Information Design B

Panel X	Stage	Easy Route	Moderate Route
	1*	Module (M)	Module (M)
	2	Module (D)	Module (D)
	3	Module (D)	Module (D)
	4	Module (D)	Module (D)
Precision at Passing (-1.0)			

*Note the stage 1 module is the 'routing' module and is the same for each candidate assigned to this panel. A total of 9 different modules are used for one panel.

Talk about item enemies here…

Discussion

References

Blackmore, D (XXXX) …….

Breithaupt, K., & Hare, D.R. (2007). Automated Simultaneous Assembly of Multistage Testlets for a High-Stakes Licensing Examination. Educational and Psychological Measurement, 67 (1) 5-20.

Hambleton, R.K. & Swaminathan, H. (1985). *Item Response Theory: Principles and Applications.* Kluwar Academic Publishing: Norwell MA.