
Xinrui Wang

Pearson VUE

Xiao Luo

Measured Progress
Fully Adaptive Multistage Testing

Abstract

This study introduces the fully adaptive multistage testing model (fa-MST) that coalesces the high testing efficiency of computerized adaptive testing (CAT) and the rigorous test quality control of computerized multistage testing (ca-MST). In fa-MST, multistage panels are assembled in the design phase and items are adaptively administered in the administration phase. A large-scale simulation study was conducted in the context of licensure testing, and the results suggested that with two layers of adaptivity, fa-MST could significantly reduce test length while maintaining high classification accuracy. Operationally, fa-MST allows the test to be managed holistically and collectively at the test level before administration.

*Keywords:* computerized adaptive testing, multistage testing, testing efficiency, test assembly regulations
Introduction

Computerized adaptive testing (CAT) has been widely recognized as an efficient exam mode, as it customizes exam forms according to candidates’ ability. Despite its high efficiency, CAT may not be the optimal choice for every exam program due to some practical concerns. First, the on-the-fly test assembly prevents test forms from being reviewed in advance of administration. Test assembly regulations are consequently embedded in the selection, estimation and stopping rules of the CAT system and realized during administration. As a result, CAT comes with less assurance of test quality (largely with respect to the compliance of constraints which ensures the comparability and fairness across different test forms) than linear tests that could be thoroughly reviewed before administration. Second, CAT is normally mildly constrained in reality, because the complexity of the item selection algorithm, as well as the item pool requirements, grows exponentially along with the growth of constraints. Usually, an operational CAT is constrained in a limited number of dimensions of the utmost importance. Advanced algorithms for delivering moderately and severely constrained CAT—such as the shadow test (van der Linden, 2000)—impose high requirements on the software and hardware of the test delivery machines and have yet seen any large-scale implementations in practice. Third, since item selection depends on previous observed responses, CAT does not allow test takers to skip items or revise responses in order to prevent their gaming of the system (Wainer, 1993). Some recent research introduced modifications to the traditional CAT paradigm to allow for skipping items and change responses to some degree (Han, 2013; Wang, Fellouris & Chang, 2017).

To address these practical concerns, studies suggested using ca-MST to make tests more manageable, although it might sacrifice a small amount of measurement precision (Hendrickson,
In ca-MST, test adaptation occurs between stages and is therefore less fine-grained than CAT because each stage has only a few modules for adaptation (Jodoin, Zenisky & Hambleton, 2006; Kim & Plake, 1993; Luecht & Nungester, 1998). Although testing efficiency declines in ca-MST, the operational advantages are prominent. First, the waning adaptivity makes it possible to exhaust routing options to identify all possible test forms in ca-MST. These forms can thereafter be scrutinized concerning the constraint violations to reinforce the comparability of different test forms. This puts an additional layer of quality control on the assembled tests to minimize uncertainties in administration. Second, since test assembly is centralized in-house, ca-MST has more time and resources to engage complex constraints in test design without increasing the computing workload on test delivery devices. Lastly, because items are administered linearly within a stage, test takers are able to navigate around items and revise responses within that stage, which partly resembles the testing experience in a linear test.

Is it possible to coalesce CAT and ca-MST and create a highly efficient and regulated adaptive testing model? One attempt along this line is Zheng and Chang’s (2015) on-the-fly assembled multistage adaptive testing (OMST), which applies the CAT framework to administer ca-MST modules that are assembled on-the-fly using algorithms like shadow test (van der Linden, 2000) or weighted deviation modeling (Stocking & Swanson, 1993). In essence, it is a special case of Wainer & Kiely’s (1987) testlet CAT. A simulation study was conducted to show that OMST managed to reach the commensurate efficiency with CAT. However, OMST fails to conserve the aforementioned operational advantages of ca-MST, for it defers the test assembly to the administration phase.
Another attempt is Wang, Lin, Chang and Douglas’s (2016) hybrid computerized adaptive testing (HCAT) model, which alternates ca-MST and CAT in one test. The rationale behind this method is that CAT may “over-adapt” in early stages of a test when the measurement is rather erroneous; therefore, ca-MST is a better choice. Nevertheless, the authors suggested to limit the size of the ca-MST component to less than half of the test, meaning that more than half of the test still relies on CAT to fine-tune test forms on-the-fly for different individuals. Similar to OMST, HCAT was shown to match, or even surmount, the efficiency of CAT in a simulation study, but it remains impossible to foresee or review all final test forms before administration.

Although OMST and HCAT can maintain the efficiency of an adaptive exam, neither of them has full control over the quality of exam forms. To that end, this study proposes the fully adaptive multistage testing (fa-MST) model that unifies the psychometric advantages in CAT and the operational advantages in ca-MST. Specifically, fa-MST consists of a design phase and an administration phase. The design phase in fa-MST has little difference from that in ca-MST, which applies the similar design principles and assembly methods to build multiple parallel MST panels (e.g., Luecht & Nungester, 1998; Luo & Kim, in press). Each panel will then include routes having equivalent nonstatistical characteristics yet differentiated psychometric characteristics for different targeted subpopulations. These panels can therefore be reviewed before going into operation in order to ensure that all test forms are finely controlled and in full compliance of the test blueprint.

The major difference between fa-MST and ca-MST arises in the administration phase, where items are linearly administered in ca-MST but adaptively administered in fa-MST. This allows for a quick decrement of measurement error and possibly an early termination of the test in fa-MST. This within-panel test adaptation is implemented by the means of an individual-
specific interim item pool. At the outset of each fa-MST administration, an empty interim item pool is initiated. While the item pool is static in CAT, this interim item pool is dynamic in fa-MST: when a routing decision is made, items in the routed module are added to the interim pool. Meanwhile, items in the interim pool are sequentially administered through the “selection-estimation” cycles as in CAT until the pool is exhausted or the termination criteria are satisfied.

At a higher level, the administration phase in fa-MST contains two layers of adaptivity: stage-level and item-level adaptivity. The stage-level adaptation causes the interim pool to absorb items on a permitted route of the MST panel, and the route is decided on the basis of an individual’s real-time test performance. The item-level adaptation releases items from the interim pool based on their information. Because routes are controlled to be equivalent with respect to nonstatistical test requirements in the design phase, a simple maximum-information item selection rule suffices for the item-level adaptation in fa-MST. This parsimonious selection rule could significantly reduce the amount of live computations for delivering severely constrained adaptive tests, compared to algorithms like the shadow test and weighted deviation modeling.

The “CAT-flavored” item-level adaptivity allows to adopt another feature in fa-MST: the early routing decisions. Luo, Kim and Dickison (in press) described a method for projecting the lower-bound (LB) and upper-bound (UB) of the final \( \theta \) estimate in the CAT environment, using the projections of the items and responses in the to-be-administered section of the test. Their method can be much simplified in fa-MST because all items in the interim pool will be administered in fa-MST and there is no need to project future items. As a result, the LB and UB of the \( \theta \) estimate at the hypothetical completion of the interim pool can be estimated by merely projecting all-incorrect and all-correct responses to the remaining items in the interim pool.
Consider a 1-3-3 multistage design with routing points of -0.44 and 0.44 at Stage 1 and 2 for a fa-MST example. The test administration starts with an empty interim item pool. Stage 1 items are first released into the interim pool and administered adaptively following maximum information algorithm. After administering each item, the candidate’s ability is estimated, along with the lower- ($\hat{\theta}_L$) and upper-bound ($\hat{\theta}_U$) of the ability estimates computed by projecting all incorrect or correct responses to the remaining items in the interim pool respectively. $\hat{\theta}_L$ and $\hat{\theta}_U$ are compared against the routing points. When they both point to the same module of Stage 2 (e.g. $\hat{\theta}_L > 0.44$ and $\hat{\theta}_U > 0.44$), that module can be selected before finishing all items in the interim pool. Items in that module are released to the interim pool for continuous CAT administration. The same procedure is taken to release Stage 3 items to the interim pool. After items of all stages are released, $\hat{\theta}_L$ and $\hat{\theta}_U$ are further used, in conjunction with the 95% confidence interval (CI) rule, to make early termination decisions for the test. Specifically, the test stops when (1) the projected range of the final ability estimate excludes the cut score, (2) the 95% CI excludes cut score, or (3) no item remains in the interim pool.

Figure 1 illustrates the procedures of a fa-MST administration.
To summarize, fa-MST assembles highly regulated panels in the design phase and delivers the panel efficiently with two layers of adaptivity in the administration phase. A simulation study was conducted to evaluate the performance of fa-MST in relation to CAT and ca-MST. The simulation study was conducted in the context of high-stakes licensure testing and its results were evaluated and discussed in relation to CAT and ca-MST pertaining to measurement precision and testing efficiency.

Methods

Conditions

Three factors that are essential to a MST design were manipulated in this study: multistage configuration, test length, and item partition strategy. A total of six multistage configurations were included: 1-2, 1-3, 1-4, 1-2-2, 1-2-3, and 1-3-3. Although not exhaustive, these configurations cover a large ground of the practice. Three levels of test length were considered: short (30 items), medium (60 items), and long (90 items). And three levels of item partition strategy were compared: equal priority that divides items equally across stages, first-stage (FS) priority that gives more items to the first stage, and last-stage (LS) priority that gives more items to the last stage. These three factors were fully crossed to generate 54 conditions.

To enhance the generalizability of the simulation study, the test in each condition was moderately constrained in two dimensions: content distribution and speededness. The choice of one discrete variable and one continuous variable was also a consideration of the generalizability. One MST panel was assembled under each condition for simplicity. Multiple panels are necessary for controlling the item exposure in practice, but not necessary for the purpose of this study. The item exposure control was considered as a rather practical issue that is
heavily dependent on the context and the condition of the item pool, and therefore was not extensively investigated in this study. Each assembled panel was subsequently administered using fa-MST and ca-MST.

To establish a performance benchmark of CAT, a simulation of shadow-test CAT (ST-CAT) was conducted in each test length condition because the other two factors were irrelevant to CAT. ST-CAT was chosen for its ability of dealing with complex constraints, as He, Diao and Hauser (2014) found for severely constrained tests ST-CAT outperformed other algorithms like the weighted deviation model, the weighted penalty model, the maximum priority index.

Furthermore, each simulation was replicated 30 times for each of the 61 true θ from -3.0 to 3.0 in increments of 0.1. The cutoff point for the pass-fail classification was set to θ=0. For the purpose of this study, tests were initially allowed to continue to the full length in simulations, and the stopping rule was applied to “retrofit” the termination decisions afterwards. This allows for a comparison of results in both the fixed-length and the variable-length situations, using the same response data for any overlapping test segments.

**Item Pool**

A 600-item pool was generated using the 3-parameter-logistic (3PL) model (Birnbaum, 1968):

\[
P(\theta) = c + \frac{1 - c}{1 + \exp[-1.7a(\theta - b)]} \tag{1}
\]

where θ is the ability parameter of the test taker drawn from \(N(0, 1)\), and a, b, and c are the discrimination, difficulty, and pseudo-guessing parameters of the item drawn from \(\text{Lognormal}(0, 0.2)\), \(N(0, 1)\), \(\text{Beta}(5, 46)\) respectively. The Fisher’s information of item \(i\) is given by:

\[
I_i(\theta) = \frac{[P_i'(\theta)]^2}{P_i(\theta)Q_i(\theta)} \tag{2}
\]
where $Q_i = 1 - P_i$, and $P_i'$ is the first derivative of $P_i$ with respect to $\theta$. Two additional variables were generated for each item: a content code drawn from $[1, 2, 3, 4]$ and an expected response time drawn from a Lognormal $(4.1, 0.3)$ (parameters were based on an operational CAT). Table 1 presents the descriptive statistics of the generated item pool, and Figure 2 shows the averaged information functions of the entire pool (solid line) and each content area (dashed lines). Each full-length test was required to have a balanced content distribution as [.4, .3, .2, .1] (set uneven to elevate the constraint severity) and an average response time of 55 to 65 seconds per item. These constraints were controlled in the design phase in MST but the administration phase in CAT.

Table 1. Descriptive Statistics of the Generated Item Pool

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Mean</th>
<th>SD</th>
<th>Min.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>a (discrimination)</td>
<td>1.02</td>
<td>0.21</td>
<td>0.51</td>
<td>1.76</td>
</tr>
<tr>
<td>b (difficulty)</td>
<td>0.06</td>
<td>1.04</td>
<td>-2.81</td>
<td>3.28</td>
</tr>
<tr>
<td>c (pseudo-guessing)</td>
<td>0.11</td>
<td>0.05</td>
<td>0.01</td>
<td>0.29</td>
</tr>
<tr>
<td>Response Time (in seconds)</td>
<td>62.44</td>
<td>17.61</td>
<td>23.95</td>
<td>131.89</td>
</tr>
</tbody>
</table>

Figure 2. The averaged information functions of the entire pool and each content area

*Multistage Tests*
The MST panels were assembled using the top-down approach (Luo & Kim, in press). First, the testing population—presumably $\theta \sim N(0, 1)$—was divided into $k$ equal-size subpopulations, where $k$ is the number of modules in the final stage. Second, these subpopulations were mapped with multistage routes, and the objective of each route was to maximize the information over its targeted subpopulation. Third, constraints on test length, content distribution, and speededness were imposed on each route. Fourth, the stage-size constraints were added to enforce the item partition conditions. For instance, the equal priority partition assigned items as $[1/2, 1/2]$ for 2-stage MSTs and $[1/3, 1/3, 1/3]$ for 3-stage MSTs, the FS priority partition assigned items as $[2/3, 1/3]$ and $[3/6, 2/6, 1/6]$, and the LS priority partition assigned items as $[1/3, 2/3]$ and $[1/6, 2/6, 3/6]$.

Panels were assembled in R (R Core Team, 2016) using the xxIRT package (Luo, 2017) using an open source mixed integer linear programming solver (lp_solve 5.5) with a time limit of five minutes. Constraints were all satisfied in assembled panels. Figure 3 presents the route information functions (RIFs) of the panels assembled under the 90-item conditions, which are the most challenging assembly conditions due to the number of required items. These RIFs indicated decent comparability among homogeneous routes (routes targeting the same subpopulation) and idiosyncrasy among heterogeneous routes (routes targeting different subpopulations). Overall, they provided a good amount of information over a wide range of the $\theta$ scale. The assembly outcomes were congruent with the design intents. Because the same pattern repeated in the 30- and 60-item conditions, results were omitted for those conditions.
The administration of a MST panel was also simulated in R. Fixed $\theta$ points were used in the distribution-based routing rule to evenly distribute the population to modules in a stage—for example, $\theta=0$ for a 2-module stage, $\theta=[-0.44, 0.44]$ for a 3-module stage, and $\theta=[-0.67, 0.00, 0.67]$ for a 4-module stage. An individual’s $\theta$ was estimated using MLE, with limits of $\pm 3.5$. A test was terminated in fa-MST when meeting three conditions: (i) it has administered at least 1/3 of items, (ii) it is in the final stage, and (iii) the cutoff point is outside of the 95% confidence interval (CI) or the projected bounds of the $\theta$ estimate.

**Shadow-test CAT**

The ST-CAT algorithm assembles a shadow test and selects the most informative item from the shadow test at each time of item selection. A shadow test is a test that fills the unfinished part of the ongoing test with items from the pool to meet all nonstatistical constraints.
and maximize the information at the current $\theta$ estimate. Shadow tests were assembled in R using the same framework and tools that assemble MST panels. The same $\theta$ estimation method was applied in ST-CAT as in fa-MST. A test was terminated when meeting two conditions: (i) it has administered at least $1/3$ of items, and (ii) the cutoff point is outside of the 95% CI of the $\theta$ estimate. The projection-based stopping rule, as well as other recent innovations, was not applied in the ST-CAT simulator because a general CAT setup is deemed to have greater generalizability than a fine-tuned one.

Results

*Measurement Precision in Fixed-length Situation*

The squared errors between the true and estimated $\theta$s were computed for each full-length simulation replication (without early terminations). Averaging the squared errors over the $\theta$ scale and replications yielded 162 (3 models by 54 conditions) mean square errors (MSEs) ranged from 0.03 to 0.24 with a mean of 0.08, and their standard errors (SEs) ranged from 0.01 to 0.08 with a mean of 0.03. The small SEs justified the chosen number of replications. The ANOVA of MSEs suggested a significant difference among models [$F(2, 159) = 41.0, p < .01$] and test lengths [$F(2, 159) = 59.5, p < .01$], but not among multistage configurations [$F(5, 156) = 1.5, p = .20$] and partition strategies [$F(2, 159) = 0.9, p = .39$]. Comparing models, ST-CAT produced lower MSE ($M = 0.04, SD = 0.01$) than fa-MST ($M = 0.11, SD = 0.05$) and ca-MST ($M = 0.11, SD = 0.05$). Comparing test lengths, the MSE was lowest in the 90-item condition ($M = 0.05, SD = 0.02$), followed by the 60-item ($M = 0.07, SD = 0.03$) and 30-item condition ($M = 0.13, SD = 0.06$).
Figure 4 presents MSEs of the 30-item conditions in more details. Although the ANOVA suggested a significant difference between CAT and MST, the difference mainly arises on the outskirts of the $\theta$ scale rather than the center of the scale. Such observation is unsurprising, given that MSTs often choose to prioritize the measurement precision in more populous and critical regions in the design. The averaged MSE of the white area of each subplot which includes 90% of the population (from -1.64 to 1.64) was 0.04 for fa-MST, 0.04 for ca-MST, and 0.03 for ST-CAT, and the averaged MSE of the gray areas was 0.18 for fa-MST, 0.19 for ca-MST, and 0.05 for ST-CAT.
Figure 5. The population-level classification accuracy at the full length

Figure 5 summarizes the population-level classification accuracy of the full-length tests. To approximate the classification accuracy for a $N(0, 1)$ population, the classification accuracies were summed across the $\theta$ scale weighted by the normalized density of each $\theta$ where the simulation was conducted. They ranged from .90 to .96 in all conditions. The classification accuracy, averaged over conditions, was 0.94 for fa-MST, 0.94 for ca-MST, and 0.95 for ST-CAT.

Efficiency and Precision in Variable-length Situation

The item-level adaptivity allows fa-MST to have variable length, and Figure 6 compares the final test length between fa-MST and ST-CAT in the 30-item conditions (lines were
smoothed using the LOESS method [Cleveland & Devlin, 1988] to enhance the visualization. ST-CAT delivered an expected pattern, where more items were used in the region around the cutoff point and fewer items in regions leaving the cutoff point. The fa-MST model delivered a similar pattern, implying the effectiveness of the early termination mechanism. Compared with the fixed test length in ca-MST (30 items in this case), fa-MST significantly reduced the test length in the vicinity of the cutoff point. Compared with ST-CAT, fa-MST tended to use even fewer items around the cutoff point but more in other regions. The former might be attributed to the application of the θ projection in the stopping rule, and the latter might be attributed to the reduced item availability and test adaptivity of the interim pool in fa-MST as compared to that of the entire item pool in ST-CAT.

To better understand the impact on the test length for the whole population, the population-level test lengths were approximated using the same method of aggregating the classification accuracies. Table 2 summarizes the population-level test length for each factor, where Δ is the difference between fa-MST and ST-CAT. Of all configurations, the simplest 1-2 MST used least number of items—nearly on a par with CAT, and the test grew longer as the configuration became more complicated. Of all partition strategies, the LS priority strategy yielded the most efficient performance, which is also very comparable with CAT. In all test length conditions, fa-MST used significantly fewer items than the maximum length. Specifically, it reduced the length by 41.3%, 43.7%, and 44.4% in the 30-, 60-, and 90-item conditions, but still used 8.0%, 14.9%, and 20.9% more items than ST-CAT in those conditions.
Figure 6. The final test length in the 30-item conditions

Table 2. The Population-level Test Lengths Between fa-MST and ST-CAT

<table>
<thead>
<tr>
<th>MST Configurations</th>
<th>Test Length</th>
<th>Partition Strategy</th>
<th>Maximum Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>MST</td>
<td>CAT</td>
<td>Δ</td>
<td>MST</td>
</tr>
<tr>
<td>1-2</td>
<td>28.5</td>
<td>29.0</td>
<td>-0.6</td>
</tr>
<tr>
<td>1-3</td>
<td>33.0</td>
<td>29.0</td>
<td>4.0</td>
</tr>
<tr>
<td>1-4</td>
<td>33.0</td>
<td>29.0</td>
<td>4.0</td>
</tr>
<tr>
<td>1-2-2</td>
<td>33.8</td>
<td>29.0</td>
<td>4.7</td>
</tr>
<tr>
<td>1-2-3</td>
<td>36.9</td>
<td>29.0</td>
<td>7.9</td>
</tr>
<tr>
<td>1-3-3</td>
<td>37.7</td>
<td>29.0</td>
<td>8.7</td>
</tr>
</tbody>
</table>

Table 3 compares population-level classification accuracy between fa-MST and ST-CAT.

The accuracy in fa-MST ranged from .90 to .97. Compared with CAT, the difference ranged
from -0.03 to 0.02, indicating a very comparable performance between fa-MST and ST-CAT in terms of classification accuracy.

Table 3. Population-level Classification Accuracy between fa-MST and ST-CAT

<table>
<thead>
<tr>
<th>Maximum Length</th>
<th>Partition Strategy</th>
<th>MST Configurations</th>
<th>ST-CAT</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>Equal Priority</td>
<td>.92 .93 .90 .93 .95 .93 .93</td>
<td></td>
</tr>
<tr>
<td></td>
<td>FS Priority</td>
<td>.92 .91 .90 .92 .92 .92 .93</td>
<td></td>
</tr>
<tr>
<td></td>
<td>LS Priority</td>
<td>.94 .92 .93 .93 .93 .94 .93</td>
<td></td>
</tr>
<tr>
<td>60</td>
<td>Equal Priority</td>
<td>.96 .95 .96 .94 .95 .95 .96</td>
<td></td>
</tr>
<tr>
<td></td>
<td>FS Priority</td>
<td>.94 .95 .94 .94 .95 .95 .96</td>
<td></td>
</tr>
<tr>
<td></td>
<td>LS Priority</td>
<td>.93 .94 .95 .94 .95 .96 .96</td>
<td></td>
</tr>
<tr>
<td>90</td>
<td>Equal Priority</td>
<td>.95 .96 .95 .95 .96 .96 .96</td>
<td></td>
</tr>
<tr>
<td></td>
<td>FS Priority</td>
<td>.95 .95 .97 .95 .94 .96 .96</td>
<td></td>
</tr>
<tr>
<td></td>
<td>LS Priority</td>
<td>.95 .95 .95 .95 .96 .95 .96</td>
<td></td>
</tr>
</tbody>
</table>

Discussion and Conclusions

This study proposed the fully adaptive multistage testing model that merges CAT and ca-MST to create a highly efficient and regulated model for delivering large-scale adaptive tests. In fa-MST, multistage panels are assembled under full control in the design phase and are adaptively administered in the administration phase. A large-scale simulation study was conducted in the context of licensure testing to compare fa-MST with ca-MST and CAT. The results indicate that fa-MST could significantly shorten the test without compromising the classification accuracy. This is a great improvement to the testing efficiency of traditional multistage testing.

Compared with CAT, fa-MST takes a different approach to regulating the test quality. Sophisticated algorithms (e.g., shadow test, weighted deviation modelling, and maximum priority index) are often used in CAT to control nonstatistical test requirements during the administration so that comparable test forms are assembled on-the-fly. Because CAT cannot
foresee the remaining items in the test, these algorithms have to rely on heuristics to refrain item selection from being too “locally greedy” while still trying to find the global optimum of the entire test. Examining a CAT administration in retrospect, it will not be surprising to see that many locally optimal item selections may not be globally optimal in the long term, especially in the early stage of the CAT.

In contrast, fa-MST manages the overall quality of a test through the holistic MST panel design and assembly which occurs before the administration. In spite of being dynamic during the administration, the interim item pool only absorbs items on a permissible route, and routes are all controlled in the design phase to be equivalent in terms of nonstatistical constraints. This approach results in very disciplined and rigorous test forms that should all be in compliance of prescribed test requirements. Consequently, the test adaptation in fa-MST becomes a lightweight process that needs to evaluate the item information only, which is similar to an unconstrained CAT environment.

Compared with the traditional ca-MST framework, fa-MST shows its higher efficiency in the administration phase. The item-level adaptivity in the administration of fa-MST accelerates the reduction of measurement error, especially at the early stages of the test. And the fast error reduction facilitates the early routing decisions. Moreover, the item-level adaptivity makes it possible to deliver variable-length tests in fa-MST, and results indicated a significant improvement in testing efficiency. However, for the fairness and validity, the stopping rule is better to be put into effective after the test arrives at the final stage, at which point the full-length test is known. Otherwise, test takers may feel deprived of opportunities to improve their performances.
Two practical implications arising from the simulation study should be highlighted. First, fa-MST tends to perform better under simple multistage configurations than under complex configurations. Simpler configurations have fewer modules and/or fewer stages. Reducing the number of modules means that routing is easier because each module accommodates a more diverse subpopulation with a broader range of ability. Reducing the number of stages means that each module includes more items that enriches the adaptivity of the interim pool more than a small-size module. Although reducing the number of stages and modules restricts the stage-level adaptivity, the simulation study indicated that the advantages of a simple configuration outweighed the disadvantages.

Second, the item partition strategy also matters. Although the FS priority partition increases the adaptivity of the interim pool in earlier stages, it delays the routing decisions because the projection of the θ estimate grows wider with too many forecasted responses. The simulation study showed that the early routing advantages in the LS-priority partition strategy was more essential to the overall testing efficiency than the initial adaptivity advantage in the FS-priority partition strategy.

This study has its limitations by design. First, the item exposure control was not extensively investigated, for the interest of this study was on the testing efficiency and test regulations. Second, only two nonstatistical constraints (content distribution and speededness) were included in the multistage design. In reality, a test might have a larger quantity of constraints in more dimensions. However, these two constraints were chosen to demonstrate how to control the categorical and the quantitative constraints in fa-MST. Additional constraints can be easily regulated using the top-down assembly approach applied in this study. Including more constraints does not always improve the generalizability though, because it might compound the
test assembly outcomes with the limitations of each idiosyncratic item pool, which could overcomplicate the study. Third, because this study was in the context of licensure testing, another study is needed in the future to evaluate fa-MST in the context of measuring individual differences.
References


