An Adaptive-within-Testlet Item Selection Method with Both Testlet Level and Test Level Content Balancing in CAT

Yuehmei Chien
Chingwei David Shin

January 2013
An Item Selection Procedure for Testlets in CAT

Abstract

Testlets are items bundled together based on related context such as a common stimulus, passage, theme, or scenario. The most important essentials of testlets in computerized adaptive testing (CAT) are content balancing, which secures the test validity, and administering items adaptive to the test taker’s ability, which ensures the test reliability. For testlets, it is not uncommon that the number of items within a testlet is large and therefore only a partial set of the testlet should be administered. Therefore, the purpose of this study is to propose a heuristic item selection procedure, which selects a testlet and a subset within the selected testlet to be administered with consideration of content balancing and being adaptive within the testlet.

This proposed heuristic procedure forms appropriate subsets for each testlet beforehand, where those subsets satisfy specific constraints at the testlet level. Those subsets are used as the unit for item selection. To deal with content balancing at the test level, the concept of the shadow test is applied to assemble a test with several subsets that contains all items previously administered to the test taker and has no content constraint violation. Within the assembled shadow test, a testlet associated with the free items is selected. Based on this selected testlet, a subset of items associated with this testlet is adaptively reselected at each item selection level, which contains the items previously administered from this selected testlet and free items within this selected testlet as well. Based on the reselected subset, one free item is randomly selected to be administered.

Simulations were conducted to evaluate the performance of this proposed heuristic procedure. The evaluation criteria include bias, mean square error (MSE), and conditional standard error of measurement (CSEM) for measurement precision, and item usage rate distribution and maximum item exposure rate for pool usage. The results presented the evidence
that the proposed heuristic is an efficient algorithm regarding the measurement precision and pool usage for administering large-size testlets in CAT.
An Adaptive-within-Testlet Item Selection Method

with Both Testlet Level and Test Level Content Balancing in CAT

Introduction

With the growth of research on computerized adaptive testing (CAT), developments of CAT have been refined and are becoming increasingly popular. However, as CAT is put into practice, there are still concerns arising, such as content balancing. As van der Linden mentioned (2005b), “the two primary requirements that test forms in a standardized testing program have to meet are similar content for each test taker and reliable estimates of their abilities.”

In the literature, the first content balancing method is called Constrained CAT (CCAT; Kingsbury & Zara, 1989), which is a straightforward method of looking for items in the content category farthest below its target percentage. Besides the CCAT, more elaborate content balancing methods include the a-parameter stratification strategy with content blocking (a_STR_C; Yi & Chang, 2003; Cheng, Chang, & Yi, 2007), the Weighted Penalty Model (WPM; Shin, Chien, Way, and Swanson, 2009), the Weighted Deviations Model (WDM; Stocking & Swanson, 1993), and the shadow test approach (ST; van der Linden, 2000).

Among those approaches, the ST approach is the only one that guarantees the satisfaction of all content constraints. A shadow test is a complete test form assembled in real time prior to each item selection that 1) satisfies all constraints specifically imposed on the adaptive test, 2) has maximum information at the current ability estimate, and 3) contains all items previously administered to the test taker (van der Linden, 2000). After a shadow test is formed, an item is selected from a shadow test instead of a pool. Van der Linden (2005a, 2005b) showed that the ST approach guarantees that all constraints are met without noticeable loss of measurement precision. Obviously, the ST approach owns two critical advantages—meeting all constraints and
maintaining measurement precision; hence, the ST has been one of the most outstanding approaches in content balancing.

The implementation of the ST approach uses the integer solver (van der Linden, 2005a, 2005b; Veldkamp & van der Linden, 2008) in the commercial software package CPLEX 9.0 (ILOG, Inc., 2003). The average time for selecting an item is about 0.43 seconds for one shadow test and 1.01 seconds for two shadow tests (Veldkamp & van der Linden, 2008). In practice, the online adaptive tests need less time for selecting an item; especially during a peak season, there might be thousands of students taking the tests on the same day. Therefore, in the research, we proposed a heuristic procedure for item selection that incorporates the concept of the shadow test so that the two advantages of the ST approach can still be remained. How the proposed heuristic procedure forms the shadow test, however, is different from the ST approach given the proposed heuristic procedure does not use an integer solver. In this study, the proposed heuristic procedure is specifically designed for adaptive tests with testlets. In general, testlets are items bundled together based on related context such as a common stimulus, passage, theme, or scenario. However, the proposed heuristic procedure in this study does not limit itself to these types of testlets as it can be applied to testlets with items bundled for any reason. Moreover, the testlets applied in the study are larger in size and, in general, contain more than ten items. In practice, for a large-size testlet, only a subset of items in the testlet will be administered instead of the whole testlet. Therefore, how to adaptively select subsets of items in a testlet is another focus to be addressed in this study, which will be described in detail in a later session.

Based on the two focuses—incorporating the concept of the shadow test into the item selection and adaptively selecting items within a testlet—this study proposed a new heuristic procedure that performs the item selection for testlets in CAT. This proposed heuristic procedure
should contain the following characteristics, which are also the main steps of this heuristic procedure: 1) forming appropriate subsets for each testlet, where those subsets satisfy their own special testlet level constraints; 2) assembling shadow tests in real time with subsets not the individual items; 3) forming one shadow test, from which a testlet associated with one of the subsets in the shadow test is selected to be administered; and 4) adaptively selecting an item to be administered from that chosen testlet. The following sections present the details of the proposed heuristic procedure and a simulation study conducted to evaluate the performance.

**Heuristic Procedure**

Shadow test assembling is a typical combinatorial optimization problem that searches for the best solution from a wide variety of possible solutions. For this type of problem, the searching space is usually vast. For example, for a 100-item pool and a test consisting of 10 items, the number of different possible test forms is \( C(100,10) \) or \( 1.73103095 \times 10^{13} \). This is an immense number even though, in practice, the pool and the test length are small. Therefore, some technique must be applied to assemble shadow tests. *Branch and Bound* (BB) is by far the most widely used tool for the combinatorial optimization problem. BB is a general algorithm that consists of a systematic enumeration of all possible solutions, where large subsets of solutions impossibly optimal are discarded. For each specific problem type, the BB algorithm should be customized; that is, the detailed design of an efficient BB algorithm is different for various optimization problems. By utilizing the BB algorithm, the proposed heuristic procedure prunes the search spaces into a reasonable size by discarding the subsets of testlets that are not informative for the test takers given the test takers’ current ability estimate. Beside BB, the most critical key techniques/concepts in this proposed heuristic, which makes the assembled shadow test possibly optimal without using any integer solver, are sorting subsets of the testlets by their
average information and researching the optimal subsets after a testlet is selected to be administered. In this proposed heuristic procedure, the unit for item selection is actually the subset of the testlet not the individual item.

This proposed heuristic provides a solution for selecting items from larger-size testlets (only a subset of the items within each testlet should be administered), and for selecting items adaptively within testlets while simultaneously balancing both the testlet level and test level content constraints. The design of this proposed heuristic procedure contains two different phases—the preparation phase and the test administration phase. The preparation phase prepares the subsets that satisfy the testlet level content constraints and are independently sorted by their average subset information for different theta levels. Those subsets are used as the unit for item selection in the test administration phase. The preparation phase only needs to be done once for a given pool before any test taker actually starts the test. After the preparation phase, the test administration phase includes two different tasks—testlet selection and subset/item selection from the selected testlet. The following sections describe the details for the preparation phase and the test administration phase.

**Preparation Phase**

**Testlet level Constraint**

The constraints in this research are not limited to the content categories. They can be other constraints such as enemy constraint, key distribution constraint, item format, item measuring, and so on. In general, if applicable, it is eloquent to specify content constraint not only at the overall test level but also at each individual testlet level. The following example explains why having the testlet level constraints for each testlet is meaningful. For a passage with 20 items, this highlights how to select a subset containing six items that measures four reading
aspects—forming a general understanding, developing interpretation, making connections, and examining content and structure. Ideally, the numbers of items in the subset that is measuring different aspects should be controlled, such as one item measuring the first aspect, and one or two items measuring the second, the third, and the fourth aspects, respectively. These types of constraints that are associated with each individual testlets and better controlled within testlets are called testlet level constraints in contrast with test level constraints. Basically, for a large testlet, using the testlet level constraints is necessary and has two advantages. First, the items administered from a specific testlet have similar content coverage across test takers. This is very important for testlet-based tests in terms of validity. For example, if a reading test should have 20% to 30% of items measuring “forming a general understanding” and the number of items measuring this aspect is not controlled at the testlet level, it is likely that the items measuring “forming a general understanding” in a test are all selected from the same testlet. The second advantage of using the testlet level constraints is that the possible subsets of testlets are largely pruned, which greatly benefits the shadow test assembling in the proposed heuristic.

**Assembling Subsets**

Given the testlet level constraints, the subsets of testlets that satisfy those testlet level constraints can be formed. The subset length (i.e. the number of items in a subset) is not necessary fixed across different testlets. The number of subsets that can be formed for a specific testlet can be huge, and some of them are actually not informative to all of the test takers or some of the test takers. For example, if a subset contains both extremely easy and difficult items, this subset is not useful for any test takers. Therefore, the only subsets that are formed and considered as candidates to be administered are ones that satisfy the testlet level constraints and are informative to certain ability test takers. There are five steps to form the candidate subsets for a
An Item Selection Procedure for Testlets in CAT

testlet: 1) the ability range is divided into six evenly spaced theta levels from -3 to 3, and $\theta_k$ is the mid-point for theta level $k$; 2) the items in the testlet are sorted by the item information for each different theta level; 3) $Y$ subsets are assembled for each $\theta_k$ using those sorted items in which the items with larger information are considered first and $Y$ is a fairly large number such as 500; 4) for each $\theta_k$, those subsets are sorted by their average subset information; 5) and it discards the subsets that do not satisfy the testlet level constraints or violate the upper bounds of the test level constraints. Note that the test level lower bounds should not be satisfied by one subset; therefore, it is not necessary to check the test level lower bounds for each subset in Step 5. However, if a subset violates a test level upper bound, the test contains that subset will definitely have at least one upper bound violation; therefore, that subset should be discarded during the preparation phase. For ease of reference, for any testlet $m$ and theta level $\theta_k$, the collection of informative subsets is denoted as $Sub_{m,k}$. The number of subsets in $Sub_{m,k}$ is less than 500 since some subsets are excluded in Step 5.

Test Administration Phase

As mentioned previously, the test administration phase regarding the item selection involves two different tasks—the testlet selection and the subset/item selection within the selected testlet. For testlet selection, one shadow test is first assembled. After a shadow test is formed, one most informative free testlet in the shadow test is selected and this selected testlet is set active. Within this active testlet, a subset is reselected from the active testlet to ensure that it is the optimal subset within the active testlet. For this optimal subset, one item is randomly selected and administered as the first item from this active testlet. The next items being administered from that active testlet is not necessary from the same subset as the first item since that subset may no longer be the optimal subset when the theta is re-estimated after the test taker
answers the first item from it. Therefore, a subset is reselected from the active testlet, which should be the most informative one given the updated theta and contain the items previously administered from this active testlet. Thus, the item selection is adaptive within the testlet. The details of this phase are described in the following sections.

**Testlet selection**

In order to ensure there is no violation of test level constraints, the concept of the shadow test is applied to this proposed heuristic. As aforementioned, the shadow test is defined as a complete test form assembled in real time prior to each item selection that—1) satisfies all constraints, 2) has maximum information at the current ability estimate, and 3) contains all items previously administered to the test taker (van der Linden, 2000). To fit the testlet situation, the shadow test is redefined in this study as a complete test form assembled in real time prior to each testlet selection that 1) contains subsets from different testlets, which satisfy all testlet level and test level constraints; 2) administers optimal subsets within testlets at the current ability estimate; and 3) contains all items previously administered to the test taker.

After a shadow test is assembled with the subsets from different testlets, one free testlet that is associated with the most informative subsets in the shadow test is selected. To describe the shadow test assembling process in detail, assume that \( j - 1 \) subsets associated with \( j - 1 \) different testlets have been administered, where \( j \) is an integer greater than or equal to 1. When \( j \) equals to 1, it is the first testlet selection. Also, assume that \( b \) subsets from \( b \) different testlets are needed for the rest of the test, and the current ability estimate \( \hat{\theta} \) is corresponding to the theta level, \( \theta_k \).

In the process of assembling a shadow test, one subset is first picked from one free testlet and then the second subset is picked from another free testlet and so on. Since there are \( b \) subsets
needed to be selected, for ease of reference, each subset selection stage is called \( \text{stage}_g \), where \( g \) = 1 to \( b \). The following are detailed steps to select the subsets for the \( \text{stage}_1 \) to \( \text{stage}_5 \), which form a shadow test with those previously administered items.

A. For \( \text{stage}_1 \), obtain the first \( N_1 \) (an arbitrary large number such as 200) subsets from each free \( \text{Sub}_{m,k} \), where \( \text{Sub}_{m,k} \) are a group of sorted informative subsets from free testlet \( m \) given the theta level \( \theta_k \). Note that the previously administered testlets are blocked from further administration in the same test, and those testlets or subsets in those testlets that are not administered and not enemies against the previously administered testlets are referred to as free testlets or free subsets. (The enemies are the items that should not administered to the same test taker for various reasons such as one providing hint to another or sharing the same content in the question.) Those \( N_1 \) subsets from each free testlet are put together, which were originally sorted within testlets. After their compilation, they are sorted again by their average subset information given \( \theta_k \). Then \( N_2 \) (an arbitrary large number such as 500) subsets are obtained from the top of the sorted subsets. Those \( N_2 \) subsets exclude the subsets that have any upper bound violation with the items previously administered. It is possible that actual number of subsets available is smaller than \( N_2 \) because many subsets are excluded. In this case, the number \( N_1 \) needs to be increased. For ease of reference, those \( N_2 \) subsets is denoted as \( \text{subsets}_j \), where \( j \) indicates those subsets are possible subsets to assemble the shadow test for the selection of the \( j \)-th testlet to be administered. Each individual subset in \( \text{subsets}_j \) is denoted as \( \text{subset}_{j,d} \), where \( d = 1 \) to \( N_2 \). This represents the
order of the subsets in $\text{subsets}_j$ based on their average subset information. For example, 

$\text{subset}_{ji}$ is the most informative subset within $\text{subsets}_j$.

B. The variables $f_1$ to $f_h$ are used to flag the position of the possible candidate subset in 

$\text{subsets}_j$ for $\text{stage}_1$ and $f_1$ is set 1 initially.

C. For $\text{stage}_1$, $\text{subset}_{ji}$ is temporarily selected as a subset candidate to form a shadow test from the subset in the position of $f_1$ from $\text{subsets}_j$.

D. For $\text{stage}_i$ between the $\text{stage}_2$ to $\text{stage}_{b-1}$, sequentially examine the subset starting from the $f_{i-1} + 1$ subset in $\text{subsets}_j$ until one subset is eligible. Set $f_i$ to the position of this eligible subset in $\text{subsets}_j$. A subset is eligible if it 1) is not associated with the same testlet as those candidates in $\text{stage}_1$ to $\text{stage}_{i-1}$, and 2) does not cause any upper bound violation with the items previously administered and the subset candidates from $\text{stage}_1$ to $\text{stage}_{i-1}$.

If no subset is eligible for $\text{stage}_1$, go back to the previous $\text{stage}_{i-1}$. If $i > 2$, set $f_{i-1} = f_{i-1} + 1$ and go to the previous $\text{stage}_{i-1}$; otherwise, set $f_1 = f_1 + 1$ and go to Step C. If all candidate subsets are selected for $\text{stage}_2$ to $\text{stage}_{b-1}$, move to Step E.

E. For $\text{stage}_b$, how the subset is picked is similar to Step D except one requirement is added to the criteria for being eligible. The selected subset must not cause any lower bound violation with the items previously administered and the items from $\text{stage}_1$ to $\text{stage}_{b-1}$. If this subset can be found, then a shadow test is formed; otherwise, go back to the previous $\text{stage}_{b-1}$ in Step D. If no subset can be found after exhausting all of the $N_2$ subsets, it may means that either the constraints are difficulty to meet or the numbers $N_1$ and $N_2$ are not set properly.
Since in practical a series of CAT simulations will be conducted before a CAT test goes operational, these issues should be revealed in the simulations and solutions can be planned to prevent them in the operational testing.

Those selected subsets from different stages are then ordered by their average subset information and denoted as \( s_1 \) to \( s_b \). Those \( s_1 \) to \( s_b \) with the previously administered subsets are the subsets from different testlets in the shadow test. One testlet associated with the most informative subsets \( s_j \) is selected as the \( j \)-th testlet to be administered and is denoted as \( T_j \).

**Subset Selection Given Testlet**

The testlet \( T_j \) is set active for subset selection based on the fact that it is associated with the most informative subset \( s_j \). However, this most informative subset within testlet \( T_j \) might or might not be the most informative given \( \hat{\theta} \) because the most informative subset are calculated using \( \theta_k \) to reduce the time spending on assembling shadow test. However, ideally, the subset administered should be adaptive to \( \hat{\theta} \) not \( \theta_k \). Therefore, a subset is reselected from \( T_j \) if it satisfies the following criteria:

1. it is eligible in the sense that it can replace \( s_j \) and form a shadow test with other subsets \( s_2 \) to \( s_b \) and the previously administered subsets;
2. it is the most informative subsets among those eligible subsets based on the previous criterion given \( \hat{\theta} \).

Once a subset is selected, one item is then randomly selected from that subset to be administered. The rest of item selection from the active \( T_j \) uses the same manner.

**Simulation Study**
In this study, the proposed heuristic procedure was evaluated through three different simulations. The first simulation was conducted with both testlet level and test level constraints using this proposed heuristic. The second simulation was conducted only with testlet level constraints. This is a testlet CAT without considering test level content balancing, where the testlets are assembled beforehand and those testlets associated with the same passage are enemies with one another and are blocked from being administered to the same test taker. Since those testlets are administered without test level constraints in the second simulation, it is considered as a testlet CAT without content balancing for the whole test. The third simulation was conducted with both testlet level and test level constraints but utilizing the simulated true ability values instead of $\hat{\theta}$ in the second phase.

The differences between the results of the first two simulations allowed us to evaluate the possible measurement precision loss in the first simulation caused by content balancing at the overall test level. The results between the first and the third simulations enabled us to evaluate the differences between using the estimated thetas and using the true thetas for item selection. It is generally considered that using the true theta value for item selection, which is the third simulation, is perfectly adaptive to the test taker. On the contrary, using the estimated theta for item selection, the selected items are adaptive to the estimated thetas, which are not necessarily close to the test taker’s true ability, especially, in the beginning of the test.

**Data**

Forty testlets that have eight or more items associated with each testlet were selected from a real pool calibrated in Rasch model, resulting in 804 items to be used in the simulation. The mean of the difficulty parameters was 0.5351 with the standard deviation of 0.9113. The distribution of the difficulty parameters in the pool were plotted in Figure 1. There were six
testlet level constraints, and each item in the testlet was associated with one testlet level constraint. For the test level constraints, there were three different types of constraints as listed in Table 1.

*Figure 1.* The difficulty distribution of the pool.

Table 1

*The Testlet Level Constraints*

<table>
<thead>
<tr>
<th>Lower Bound</th>
<th>Upper Bound</th>
<th>Constraint</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item Measure</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.1</td>
<td>0.3</td>
<td>measure 1</td>
</tr>
<tr>
<td>0.04</td>
<td>0.15</td>
<td>measure 2</td>
</tr>
<tr>
<td>0.04</td>
<td>0.10</td>
<td>measure 3</td>
</tr>
<tr>
<td>0.06</td>
<td>0.2</td>
<td>measure 4</td>
</tr>
<tr>
<td>0.10</td>
<td>0.25</td>
<td>measure 5</td>
</tr>
<tr>
<td>0.10</td>
<td>0.35</td>
<td>measure 6</td>
</tr>
<tr>
<td>0.06</td>
<td>0.3</td>
<td>measure 7</td>
</tr>
<tr>
<td>0.2</td>
<td>0.2</td>
<td></td>
</tr>
<tr>
<td>Functional</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td>0.2</td>
<td>Informational</td>
</tr>
<tr>
<td>0.6</td>
<td>0.6</td>
<td>Literary</td>
</tr>
<tr>
<td>Key Distribution</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.15</td>
<td>0.5</td>
<td>key A</td>
</tr>
<tr>
<td>0.15</td>
<td>0.5</td>
<td>key B</td>
</tr>
<tr>
<td>0.15</td>
<td>0.5</td>
<td>key C</td>
</tr>
<tr>
<td>0.15</td>
<td>0.5</td>
<td>key D</td>
</tr>
</tbody>
</table>
CAT design

A fixed-length CAT of 30 items, consisting of five testlets each with six items was administered to every simulee. Two different samples were generated to report different statistics criteria for each simulation. First, a conditional sample of 50,000 simulees was generated with 10,000 simulees’ theta values at 5 equally spaced theta levels from -2 to 2. Second, a normal sample of 20,000 simulees was generated from a standard normal distribution. The initial theta point to select the first item is set at -1. The results from the conditional sample were used to report bias, mean square error (MSE), and conditional standard error of measurement (CSEM) on the five theta points (-2, -1, 0, 1, and 2). The results from the normal sample were used to report the correlation between the true and estimated ability values and the pool usage.

Results

As mentioned before, there were three different CAT simulations. Those are referred to as Simulations I, II, and III. Simulation I was conducted with both testlet level and test level constraints; Simulation II was conducted only with testlet level constraints; and Simulation III was conducted with both testlet level and test level constraints but utilizing simulated true ability values for item selection in Phase II instead of utilizing the estimated ability values. The results of bias, MSE, and CSEM for Simulations I, II, and III are listed in Table 2. To visually examine the results, bias, MSE, and CSEM are separately plotted in Figures 2, 3, and 4, respectively.

All of the simulations show very small bias in the middle theta ranges and slightly larger bias at the two end points. With these small biases all under 0.07, the three simulations had comparable results in terms of bias. Table 2

The results of bias, MSE, and CSEM for Simulations I, II, and III

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Bias</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
An Item Selection Procedure for Testlets in CAT

For the MSE, the three simulations led to very similar results across five theta points.

Simulation I, which was conducted with both test level and testlet level constraints, even had slightly smaller MSE at ability points 0, 1, and 2 than those from Simulations II and III. By examining Figure 3, the MSE values for the three simulations are very similar at the middle
range of ability and slightly different at the two ends of ability. For $\theta = 2$, Simulation I generated a smaller MSE than Simulations II and III while for $\theta = -2$, Simulation II had the smallest MSE. It can be observed from Figure 1 that the pool has very few items with item difficulty around $\theta = -2$. Therefore, it is expected that the MSE is larger for $\theta = -2$ than other ability levels for each of the three simulations. Also, it is observed that, as expected, the CAT without content balancing (Simulation II) performed better regarding the MSE than the other two simulations. On the contrary, when the pool has enough items, the proposed heuristic procedure implemented in Simulation I generated comparable results comparing to Simulations II and III regarding MSE.

![Figure 3](image.png)

**Figure 3.** MSE results for Simulations I (constrained), II (not constrained), and III (perfectly adaptive).

The CSEM of the three simulations are shown in Figure 4. For the two extreme thetas, Simulation II had a slightly better performance in terms of smaller CSEM values, which is as expected. However, for the middle range of ability, these three simulations were comparable.
Based on the simulation results conducted using the normal samples, the pool usage rate distribution and the correlation between true and estimated thetas were reported in Tables 3 and 4, respectively. For reference, the average item exposure rate is 0.0373 (= 30 / 804) if items are randomly selected. For the pool usage, it was not surprising that Simulation II had the highest rate on the items never being used and the largest maximum item exposure rate compared with Simulations I and III. Without content balancing (Simulation II), the testlet selection was not constrained and only based on average testlet information; therefore, the testlet with larger information was certainly more popular than others. In terms of the larger pool usage rate and smaller maximum exposure rate, the proposed heuristic (Simulation I) performed better than the other two simulations.
Table 3

**Pool Usage Rate Distribution**

<table>
<thead>
<tr>
<th>Item Never Used</th>
<th>Simulation I</th>
<th>Simulation II</th>
<th>Simulation III</th>
</tr>
</thead>
<tbody>
<tr>
<td>IE_Rate[ 0.000 ~ 0.005]</td>
<td>0.5348</td>
<td>0.6393</td>
<td>0.5647</td>
</tr>
<tr>
<td>IE_Rate[ 0.005 ~ 0.100]</td>
<td>0.1119</td>
<td>0.0572</td>
<td>0.1032</td>
</tr>
<tr>
<td>IE_Rate[ 0.100 ~ 0.200]</td>
<td>0.2090</td>
<td>0.1654</td>
<td>0.1878</td>
</tr>
<tr>
<td>IE_Rate[ 0.200 ~ 0.300]</td>
<td>0.0896</td>
<td>0.0846</td>
<td>0.0858</td>
</tr>
<tr>
<td>IE_Rate[ 0.300 ~ 0.400]</td>
<td>0.0361</td>
<td>0.0249</td>
<td>0.0410</td>
</tr>
<tr>
<td>IE_Rate[ 0.400 ~ 0.500]</td>
<td>0.0112</td>
<td>0.0199</td>
<td>0.0075</td>
</tr>
<tr>
<td>IE_Rate[ 0.500 ~ 0.600]</td>
<td>0.0012</td>
<td>0.0050</td>
<td>0.0062</td>
</tr>
<tr>
<td>IE_Rate[ 0.600 ~ 0.700]</td>
<td>0.0025</td>
<td>0.0012</td>
<td>0.0012</td>
</tr>
<tr>
<td>IE_Rate[ 0.700 ~ 1.000]</td>
<td>0.0012</td>
<td>0.0012</td>
<td>0.0012</td>
</tr>
<tr>
<td>Maximum IE Rate</td>
<td>0.7013</td>
<td>0.9420</td>
<td>0.7464</td>
</tr>
</tbody>
</table>

The correlation between the true theta and the estimated theta is listed in Table 4. Across the three simulations, their correlation values were extremely close to each other. Therefore, these three simulations had no difference on the correlation between the true and the estimated thetas.

Table 4

**Correlation between True and Estimated thetas**

<table>
<thead>
<tr>
<th>Simulation I</th>
<th>Simulation II</th>
<th>Simulation III</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.9334</td>
<td>0.9338</td>
<td>0.9328</td>
</tr>
</tbody>
</table>

Besides the results of measurement precision and pool usage, the time for selecting an item was ranged from 0.008 to 0.04 seconds based on Simulation I that was conducted on a personal laptop. It should be noted that the simulation program conducted for this research had not yet been tuned and optimized to speed up the item selection processing time. Therefore, it is
possible to be further improved in the future. However, based on the results of 0.008 to 0.04 seconds, it is acceptable if the proposed heuristic procedure is used on a real operational CAT.

Based on the simulation results, the conclusions are as follows. First, the proposed heuristic procedure presented by Simulation I performed well and did not have precision loss compared with Simulation II that did not control the content balancing. Second, in general, the proposed heuristic was comparable in terms of measurement precision to Simulation III that was perfectly adaptive by using the true ability for item/testlet selection. Third, the proposed heuristic performed better in item exposure control in terms of larger pool usage rate and smaller maximum item exposure rate. In short, the simulation results provided evidence that the proposed heuristic is an efficient algorithm regarding the measurement precision, pool usage, and item selection time for administering large-size testlets with both testlet level constraints and test level constraints in CAT.

**Discussion and Limitation**

The proposed heuristic of CAT item selection procedure for testlets has the following advantages. First, each testlet can have its own specific constraints or constraints that prevent selecting items within a testlet that are largely associated with a certain item property. One extreme instance is that a subset administered only measures the vocabulary skill. Second, this proposed heuristic procedure selects items that have no constraint violation as long as feasible solutions exist. Third, it is tenable for operational use given that average time for selecting an item for the simulations conducted in this study ranges from 0.008 to 0.04 seconds on a personal laptop.

One strategy used by this proposal heuristic to avoid spending too much time on finding a feasible solution in assembling a shadow test is to limit its search space by setting a reasonable
value for $N_2$, which is 25 in this study. If no feasible solution can be found from those $N_2$ subsets for a shadow test, the proposed heuristic will use the shadow test formed from the previous testlet selection phase. By doing it this way, the proposed heuristic procedure is always able to produce a shadow test if the item pool admits of at least one shadow test to be formed with those $N_2$ subsets for the first testlet selection. From another point of view, if a pool does not have enough feasible solutions to allow different shadow tests to be formed for different ability test takers from those $N_2$ subsets, it is necessary to examine the reasonableness and practicability of those testlet level and test level constraints.

Van der Linden (2005a) claimed that the ST approach is able to administer the optimal test. The test is optimal in the sense that the administered shadow test is most informative given the test taker’s current ability estimate at each item selection phase. The proposed heuristic procedure in this paper is not able to have the claim that it is able to administer the optimal test as the ST approach does since once a shadow test is found, the searching is stopped. The shadow test is also not guaranteed to be the most informative even though the most informative subsets are first used to form the shadow test. However, we are able to claim that the optimal subsets are administered given the selected testlets. One technique that the proposed heuristic procedure utilizes to achieve the optimal subset is to re-select the subset after one testlet is chosen to be administered in the testlet selection phase. This testlet is selected and set active because one of its subsets is in the shadow test, which is denoted as $subset'_{td}$. Therefore, for this active testlet, it at least has $subset'_{td}$ that is satisfied the test level constraints with other subsets in that shadow test. In order to select the optimal subset to be administered, the proposed heuristic will try to search another subset that is: 1) associated with the active testlet, 2) more informative than
An Item Selection Procedure for Testlets in CAT

$\text{subset}^t_{i,d}$, and 3) satisfying test level constraints if $\text{subset}^t_{i,d}$ is replaced by it in the shadow test.

The research results show that Simulation I is comparable to Simulation II, where Simulation I is the proposed heuristic and Simulation II is the testlet CAT without content balancing and solely selecting testlets by their informative elements. Therefore, based on the finding for this specific pool used in this study, we can claim that there is no precision loss introduced by this proposed heuristic, which administer the optimal subsets.

**Future Work**

This is an initial experimental study for a heuristic procedure that is developed for CAT item selection with testlets, in which the test is not only adaptive between testlets but also adaptive within testlets. As results have shown, this heuristic procedure is successful. However, there is only one testlet pool used in this study. In the future, it is necessary to apply this heuristic procedure to more CAT with testlets to generalize the findings.

Since this is an initial experimental study, item exposure control is not the focus of this study. With the success of this initial study, it is truly beneficial to apply more sophisticated item exposure control strategies into the heuristic procedure to increase the item usage and to better control the maximum item exposure rate.
References


