The Variable-Length Adaptive Diagnostic Testing

NCME
Chicago, Illinois

Yuehmei Chien (Pearson)
Chingwei David Shin (Pearson)
Ning Yan (Independent Consultant)

April 2015
Introduction

Recently, the diagnostic assessment, which uses the diagnostic classification models (DCMs) to determine mastery or non-mastery of a set of attributes and to provide strengths and weaknesses, has drawn much attention of the practitioners. The diagnostic assessment is adaptive to a pool of items that are specifically designated as diagnostic. Moreover, the variable-length adaptive diagnostic assessment is desirable because, when a mastery status for an attribute or a profile classification is sufficiently certain, there is no need to administer more items for that attribute or for the test; therefore, it is a very efficient tool for educators to obtain timely learning outcomes while not exhausting students with test-taking.

The goal of this study was to evaluate different adaptive algorithms in the variable-length adaptive diagnostic testing. Two new heuristics were proposed and used as part of the algorithms.

Adaptive Diagnostic Testing

The most critical components in adaptive diagnostic testing are the item-selection algorithms and the termination rules. During adaptive testing, items are sequentially selected, one item at a time, based on the performance of the respondent on the previous items. Each time when the respondent has completed a new item, the posterior distribution of the attribute profile for the respondent is updated to incorporate new information provided by the response to this item. A set of termination criteria are then checked to see if the test may end at this point. If conditions for termination are not yet satisfied, the test continues until it finally meets the criteria. And it is necessary the maximum test length is included in the termination criteria to prevent the test goes too long.
Adaptive Item Selection

For the item-selection algorithms, three different approaches were adopted including the Posterior Weighted Kullback-Leibler (PWKL; Cheng, 2009), the entropy for posterior distribution, and the entropy for marginal distribution.

Notations

Let \( \pi(\alpha) \) denote the current posterior probability of the attribute profile \( \alpha \) for the respondent based on the observed responses to the test items that have already been administered. Let \( t \) denote a candidate item for the next round of the test, and let \( X(t) \) be the random variable whose value is the response to \( t \) by the respondent. The known item parameters for \( t \) are the conditional probabilities \( P(X(t) = x | \alpha) \) where \( x \) is either 0 or 1. These are the probabilities of the two possible item responses given the attribute profile \( \alpha \). The marginal probabilities for \( X(t) = x \) are then

\[
P(X(t) = x) = \sum_\alpha \pi(\alpha) \cdot P(X(t) = x | \alpha) \quad \text{for } x \in \{0,1\}
\]

Item selection by maximizing the PWKL score

The Kullback-Leibler (KL) divergence was proposed by Kullback and Leibler in 1951. It is a measure of the difference between two probability distributions. The KL divergence \( KL_t(\hat{\alpha} || \alpha) \) defined below expresses the expected ability of the item \( t \) to distinguish between the current estimated mastery profile \( \hat{\alpha} \) and the unknown true mastery profile \( \alpha \) through the difference between the two conditional distributions \( X(t) | \hat{\alpha} \) and \( X(t) | \alpha \):

\[
KL_t(\hat{\alpha} || \alpha) = \sum_{x \in \{0,1\}} P(X(t) = x | \hat{\alpha}) \cdot \log \left( \frac{P(X(t) = x | \hat{\alpha})}{P(X(t) = x | \alpha)} \right)
\]
The value of $KL_t(\hat{\alpha} \parallel \alpha)$ is zero when $\hat{\alpha} = \alpha$ and increases as the two distributions $X(t) \mid \hat{\alpha}$ and $X(t) \mid \alpha$ diverge. Because the true $\alpha$ is unknown, a global KL score is constructed for the item $t$ as the sum of $KL_t(\hat{\alpha} \parallel \alpha)$ over all possible $\alpha$:

$$KL_t(\hat{\alpha} \parallel \alpha) = \sum_{\alpha} KL_t(\hat{\alpha} \parallel \alpha)$$

The item with maximum $KL_t(\hat{\alpha} \parallel \alpha)$ is chosen as the next item (Xu, Chang, & Douglas, 2003; Cheng, 2009). The KL score reflects the discrimination power of the item $t$ to distinguish the current estimated mastery profile from all other possible profiles. Its definition implicitly assumes that each $\alpha$ is equally likely to be the true profile for the respondent. The definition can be improved if each $\alpha$ is weighted by its current posterior probability $\pi(\alpha)$, this results in the PWKL score as proposed by Cheng (2009). The equation below defines the PWKL score:

$$PWKL_t(\hat{\alpha} \parallel \alpha) = \sum_{\alpha} KL_t(\hat{\alpha} \parallel \alpha) \cdot \pi(\alpha)$$

**Item selection by minimizing the entropy for posterior distribution**

Shannon entropy is a mathematical construct introduced by Claude Shannon in a 1948 paper (add reference). It is widely used as a measure of uncertainty in a probability distribution. For a discrete probability distribution $\pi$ that takes $N$ possible values with probabilities $p_1, p_2, \ldots, p_N$, the Shannon entropy $H(\pi)$ is defined as

$$H(\pi) = -\sum_{i=1}^{N} p_i \log_2(p_i)$$

One approach for item selection is to search for a candidate item that minimizes the expected Shannon entropy of the posterior distribution of the attribute profile for the respondent (as
described by Xu, Chang, and Douglas in 2003). We next describe the steps for calculating this expected Shannon entropy for any given candidate item.

Now assume \( t \) is given to the respondent as the next item, and let \( \pi_t(\alpha) \) denote the new posterior probability of the attribute profile \( \alpha \) after the response to item \( t \) is observed. The calculation of \( \pi_t(\alpha) \) depends on both the previous posterior distribution \( \pi \) and the new item response \( X(t) \):

\[
\pi_t(\alpha \mid \pi, X(t) = x) = \frac{\pi(\alpha) \cdot P(X(t) = x \mid \alpha)}{P(X(t) = x)} \quad \text{for } x \in \{0,1\}
\]

The Shannon entropy formula is then used to calculate the entropy of the new posterior distribution \( \pi_t \) in the case of each possible value of \( X(t) \):

\[
H(\pi_t \mid \pi, X(t) = x) = - \sum_\alpha \pi_t(\alpha \mid \pi, X(t) = x) \cdot \log_2 (\pi_t(\alpha \mid \pi, X(t) = x))
\]

The expected Shannon entropy for \( \pi_t \) is then

\[
EH(\pi_t \mid \pi) = \sum_{x \in \{0,1\}} P(X(t) = x) \cdot H(\pi_t \mid \pi, X(t) = x)
\]

An item \( t \) that minimizes \( EH(\pi_t \mid \pi) \) among all candidate items is a reasonable choice for the next round of the test. For ease of reference, this adaptive item selection approach by minimizing the entropy for posterior distribution is referred to SHE_POST thereafter.

**Item selection by minimizing the entropy for marginal distributions**

In theory the DCM is specifically designed to support and promote the multidimensional point of view, in which the central role is played by the vector-valued mastery profile that reflects the joint status on multiple attributes, and respondents are classified according to this multidimensional profile. In practice, however, there has been a persistent demand for turning
the DCM output into a collection of unidimensional statements for ease of interpretation. This is achieved by reducing the probability distribution for the mastery profile into unidimensional marginal distributions for the individual attributes. We next describe a marginalized version of the expected Shannon entropy $EH(\pi_t | \pi)$ in which the joint distribution $\pi_t$ is replaced by $\{\mu_{t,k} : k \in 1, 2, \ldots, K\}$, where $\mu_{t,k}$ is the marginal distribution of $\pi_t$ for the $k$-th attribute.

The marginal distribution $\mu_{t,k}$ is a Bernoulli distribution, which takes value 1 (code for the “mastery” status) with probability $\mu_{t,k}(1)$ and value 0 (code for the “non-mastery” status) with probability $\mu_{t,k}(0)$. It is straightforward to calculate the probability $\mu_{t,k}(s)$ for $s \in \{0, 1\}$ in the case of each possible value of $X(t)$ using the joint probabilities $\pi_t(\alpha)$:

$$\mu_{t,k}(s \mid \pi, X(t) = x) = \sum_{\alpha : \alpha_k = s} \pi_t(\alpha \mid \pi, X(t) = x)$$

where the sum is overall $\alpha$ with $\alpha_k = s$. The Shannon entropy for $\mu_{t,k}$ is:

$$H(\mu_{t,k} \mid \pi, X(t) = x) = - \sum_{s \in \{0, 1\}} \mu_{t,k}(s \mid \pi, X(t) = x) \cdot \log_2 (\mu_{t,k}(s \mid \pi, X(t) = x))$$

The expected Shannon entropy for $\mu_{t,k}$ is then

$$EH(\mu_{t,k} \mid \pi) = \sum_{x \in \{0, 1\}} P(X(t) = x) \cdot H(\mu_{t,k} \mid \pi, X(t) = x)$$

From the unidimensional point of view, it is desirable to minimize the expected Shannon entropy values for all the marginal distributions when selecting the item $t$. It is reasonable to use the min-max approach and choose the item $t$ that minimizes $max\{EH(\mu_{t,k} \mid \pi) : k = 1, 2, \ldots, K\}$. For ease of reference, this adaptive item selection approach by minimizing the entropy for marginal distributions is referred to SHE_MARG thereafter.
**Termination Rule**

In the literature, the termination rules for adaptive diagnostic testing using DCMs are largely absent. The termination rule is used when the adaptive diagnostic test is a variable length test, which makes sure the estimated profile based on the current test has reached certain measurement criteria. Beside the predefined min- and max-attribute-level test length, a statistic should be developed and used to decide when the test or the attribute should no longer be administered. In this study, several termination criteria were investigated, including the posterior probability, the marginal probability, the Shannon Entropy (SHE), and the bootstrap approach.

**The posterior probability**

Tatsukoa (2002) stated that a diagnostic assessment can be terminated when the maximum posterior probability for a class or profile (e.g., “111” is one of the eight classes for three attributes) exceeds 0.8. Using the posterior probability as the termination rule, the test is ended when the number of items administered is more than the predefined minimum test length and the posterior probability calculated based on those items exceeds the predefined value (such as 0.8 or 0.9).

**The posterior marginal probability**

Besides estimated profile, another way to obtain the mastery classification is based on the posterior marginal probabilities for attributes. Using the posterior marginal probability, the test stops when all posterior marginal probabilities for attributes either very close to 0 or 1, such as all posterior marginal probabilities either above 0.8 or below 0.2.
The Variable-Length Adaptive Diagnostic Testing

The CSEM_PWKL

PWKL is the posterior probability weighted KL information that has been used as a statistic for adaptive item selection, which selects items with larger PWKL values. Intuitively, we borrow the concept of conditional standard error of measurement (CSEM) in item response theory (IRT), the CSEM_PWKL is $1/\sqrt{\text{sum}(\text{PWKL})}$. As the test goes longer, the PWKL is getting larger and the CSEM_PWKL is getting smaller. When the CSEM_PWKL is smaller than a predefined value, the test can stop.

The bootstrap

It has been observed that the marginal posterior probability for attributes may largely increase or decrease after one more item is taken. If only applying regular termination rules, the test might be able to terminated when the marginal posterior probability happens to have a big jump, such as from 0.6 to 0.92. (add a picture?) To avoid such a condition, one might intuitively consider to have marginal posterior probability converged (such as above 0.90 for at least three consequent items) after the test length has research the minimum.

Another approach is to calculate the uncertainty of estimates using bootstrap sampling. Bootstrap sampling refers to a random sampling procedure with replacement in statistics, which provides a simple and straightforward way to derive standard errors of estimates (SEE) and confidence intervals (CI) without parametric distribution assumption on the parameters. Therefore, intuitively and naturally, bootstrap sampling can be used to access the stability of estimates of the posterior probability for attributes in DCM.

To obtain the bootstrap SEE, the steps are as follows:
• Step 1: obtain the current data set including items administered so far and their corresponding responses.

• Step 2: take the current data set as the original data set of \( n \) items and resample \( n-1 \) items from it with replacement to form a new sample of size \( n-1 \).

• Step 3: calculate the marginal posterior probability for attributes of the new sample.

• Step 4: repeat Step 2 and Step 3 for a large number of times (such as 100 or 200 iterations).

• Step 5: obtain the bootstrap SEE, that is the standard deviation of the marginal posterior probability for each attribute over those iterations.

After the bootstrap SEE for each attribute is attained, a predefined threshold can be applied to determine the termination or continuation of the test.

**Simulation**

The performance or different behavior for various adaptive item selection approaches and termination rules are examined through two simulations. In order to possibly discover the difference among various adaptive item selection approaches, the first simulation mainly focuses on adaptive item selection with a fixed test length design. The results from the first simulation include the number of attributes measured and the resulting overall and attribute-level classification accuracy. The second simulation focuses on variable test length DCM-CAT, in which two item-selection approaches and four termination rules were investigated. The simulation results were compared for overall profile classification accuracy, attribute-level classification accuracy, and test length.
Simulation I

Data Generation and Adaptive Test Design

This simulation study was conducted to compare the three item-selection algorithms: PWKL, SHE_POST, and SHE_MARG. In order to know how much gain from adaptively selecting items, the “RANDOM” item selection was included as the baseline for comparison. The test length is fixed to 20. The other components of simulation are described below.

Generated Pool

A pool of two hundreds items was generated. The slip and guessing parameters were randomly drawn from a uniform [0.15, 0.35] distribution and a uniform [0.20, 0.25] distribution, respectively. A Q-matrix with four attributes was also generated. One to two attributes were randomly assigned to each item, which results roughly half of items measuring one attribute and another half measuring two attributes.

Generated True latent Profiles

The true latent profiles were generated using a method used by Finkelman, Kim, & Roussos (2009). The latent attributes were assumed to have a multivariate standard normal distribution and then compared to specific cut points in order to determine mastery or non-mastery status for each attribute. The paired correlation of attributes was set to 0.6 and the mastery rates were all set to 0.5 for those four attributes.

Estimation

The initial profile is “0000” for selecting the first item. Then the maximum likelihood estimation (MLE) method is used to estimate the profile.
**Attribute balancing**

An attribute balancing method named Quota is adopted for balancing attribute coverage (see Chien, 2015 for more details). The Quota method selects the next item from all eligible items, in which eligible means all the constraints associated with the items are currently below the predefined upper bound of their target administration rate. The lower bound and upper bound is 25% and 50% for each attribute.

**Results**

As previously mentioned, the first simulation study was conducted to compare the three item-selection algorithms: PWKL, SHE_POST, and SHE_MARG with “RANDOM” serving as a baseline for comparison. The TABLE 1 shows the results of overall and attribute-level classification accuracy. The SHE_POST and PWKL performed best while the SHE_MARG approach is slightly inferior and the RANDOM approach is the worst. These findings are not exactly the same as those from Cheng (2009), where her study showed the PWKL approach consistently performs slightly better than SHE_POST. The results clearly show one possible benefit of using adaptive item selection for diagnostic assessment when we compare the three adaptive item selection approaches with the RANDOM method.
TABLE 1

*Overall and Attribute-Level Classification Accuracy*

<table>
<thead>
<tr>
<th>Classification Accuracy</th>
<th>profile</th>
<th>attr1</th>
<th>attr2</th>
<th>attr3</th>
<th>attr4</th>
</tr>
</thead>
<tbody>
<tr>
<td>SHE_POST</td>
<td>0.855</td>
<td>0.954</td>
<td>0.966</td>
<td>0.944</td>
<td>0.961</td>
</tr>
<tr>
<td>SHE_MARG</td>
<td>0.831</td>
<td>0.945</td>
<td>0.952</td>
<td>0.942</td>
<td>0.950</td>
</tr>
<tr>
<td>PWKL</td>
<td>0.859</td>
<td>0.957</td>
<td>0.959</td>
<td>0.950</td>
<td>0.959</td>
</tr>
<tr>
<td>RANDOM</td>
<td>0.600</td>
<td>0.874</td>
<td>0.868</td>
<td>0.846</td>
<td>0.858</td>
</tr>
</tbody>
</table>

TABLE 2 shows the average test length of the four attributes. The three adaptive item selection approaches consistently have smaller test lengths than the RANDOM method, except for the SHE_MARG method on Attribute 1. This observation indicates that overall, the three adaptive item selection approaches tend to give a test with more simple items than the one with items randomly selected from the pool. Among the three approaches, the SHE_MARG approach consistently has larger numbers of average test length across attributes, which indicates that this method selects more 2-attribute items than the other two adaptive approaches.

TABLE 2

*Average Test Lengths of Attributes*

<table>
<thead>
<tr>
<th></th>
<th>total</th>
<th>attr1</th>
<th>attr2</th>
<th>attr3</th>
<th>attr4</th>
</tr>
</thead>
<tbody>
<tr>
<td>HE_POST</td>
<td>20</td>
<td>7.583</td>
<td>7.648</td>
<td>7.239</td>
<td>7.403</td>
</tr>
<tr>
<td>SHE_MARG</td>
<td>20</td>
<td>8.280</td>
<td>8.149</td>
<td>7.601</td>
<td>7.800</td>
</tr>
<tr>
<td>PWKL</td>
<td>20</td>
<td>7.391</td>
<td>7.509</td>
<td>7.054</td>
<td>7.237</td>
</tr>
<tr>
<td>RANDOM</td>
<td>20</td>
<td>8.415</td>
<td>8.248</td>
<td>8.007</td>
<td>8.134</td>
</tr>
</tbody>
</table>
Based on what we have observed regarding the difference on the average attribute-level test length, it would be useful to examine the average numbers of attributes conditional on the sixteen true profiles. (Note that the number of simulees in each true profile is varied.) FIGURE 1 shows the results. The conditional results have the following findings: 1) compared to the method itself across different true profiles, the three adaptive approaches tend to have more simple attribute items for the low true-profiles (the number of attributes mastered is 0 or 1), and have more two-attribute items for the high true-profiles (the number of attributes mastered is 3 or 4) while the RANDOM method consistently has the same average attribute-level test length; 2) the SHE_POST and the PWKL approaches have similar patterns across different true-profiles, in which the difference between these two methods are slightly more obvious for the high true-profiles; 3) the SHE_MARG and PWKL approaches have most single attribute items for the low profiles, especially if the average attribute-level test lengths of these two approaches are both below 1.1 for the true-profile “0000”, which means the test given to the student with true-profile “0000” only has two 2-attribute items on average.

FIGURE 1. 

The Conditional Average Numbers of Attributes.

To know the possible classification accuracy results for the three adaptive item selection approaches, we further plotted the conditional classification accuracy rates for the sixteen true
profiles as shown in FIGURE 2. The RANDOM method performs relatively much better as it moves from low true-profiles to high true-profiles. Even though for the low true-profiles, the SHE_MARG approach has larger average numbers of attributes, some of the low true-profiles -- "0000" and “0010” do have the similar classification accuracy rates compared to the other two adaptive item selection methods. This observation confirms that for very low-profile students, complex items (measuring more than one attribute) does not help improve the classification.

FIGURE 2.

*The Conditional Classification Accuracy Rate across Sixteen True Profiles.*

---

**Simulation II**

**Data Generation and Simulation Design**

The second simulation focuses on variable test length DCM-CAT. In this study, four item-selection approaches and four termination rules were investigated through a series of simulations, resulting in sixteen adaptive algorithms for diagnostic tests. The sixteen item selection algorithms were compared for overall profile classification accuracy, attribute-level classification accuracy, and test length.

The four item-selection methods are those used in Simulation I, including SHE_POST, SHE_MARG, PWKL, and RANDOM. As used in Simulation I, Simulation II uses the same data.
generation methods and the same four adaptive item-selection methods. Simulation II is a variable-test-length adaptive test design, however, and it differs from Simulation I in the following aspects:

**Termination Rule**

For a variable-test-length DCM-CAT, a termination rule is used to check if the test may end at the point that it has reached the minimum test length. If conditions for termination are not yet satisfied, the test continues until it finally meets the criteria or reaches the maximum test length. Four termination rules described previously including CSEM_PWKL, PROF_PROB, ATTR_PROB, and BOOT were included. The termination values associated with the four methods are 0.1 for CSEM_PWKL, 0.8 for PROF_PROB, 0.1 and 0.9 for ATTR_PROB, and 0.28 for BOOT.

**Generated True Latent Profiles**

Two different mastery rates for attributes are manipulated, medium (labeled as Med) and high-to-low (labeled as HL). Medium sets 50% mastery rates for all attributes while high-to-low sets 75%, 60%, 45%, and 30% mastery rates for the four attributes, respectively. Two thousand simulees were generated for each of the different mastery rates. When the data is generated based on a mastery rates, the numbers of simulees in different classes differ. In order to examine the impact of different combination of item selection methods and termination rules to students in different true classes, 200 simulees in each class and 3200 simulees in total for the four-attributes test were also generated as the conditional sample.
The Variable-Length Adaptive Diagnostic Testing

Test Length

The maximum test length is 20 and the minimum test length is one of the two factors manipulated in Simulation II. The two minimum test lengths are 8 and 12 (labeled as Len_8 and Len_12, respectively), which represents the possible benefit of using adaptive diagnostic algorithm that is the test length might be shorter than for certain populations of students.

Results

There are two parts to the results. The first part is the simulation results for the sample, generated based on the mastery rates for attributes and paired correlation, which is referred to as the mastery-rate sample. The mastery-rate sample has true profiles or classes generated based on the mastery-rate and paired-correlation provided; thus, the class distribution matches our assumption about the underlying attributes relationship. The second part is the simulation results for the conditional samples, for which each class has exactly the same amount of simulees generated. The conditional samples allows us to examine the results for different types of students in different profiles while the mastery-rate sample presents the overall results of the population.

Mastery-rate Sample

Figure 3 shows the mastery-rate sample results of classification accuracy on profiles and average test length. There are four study conditions with sixteen item selection algorithms. Clearly, the RANDOM method should perform much worse than the others. For the classification accuracy on profiles, as what we expected, the larger the minimum test length, the higher classification accuracy rate across those sixteen item selection algorithms when compared Len_12 to Len_8 (row 3 to row 1 and row 7 to row 5, respectively, for the Med and HL
underlying mastery rates.) Apparently, there is some interaction between the adaptive item selection methods and the termination rules; therefore, there is no termination rule consistently performing better with a certain adaptive item selection approach across those four study conditions.

For the Med & Len_8 study condition, the termination rules ATTR_PROB and PROF_PROB with the adaptive item selection methods PWKL and SHE_POST perform better than other combinations in terms of high classification accuracy rates and shorter average test length. For the Med & Len_12 study condition, the termination rule ATTR_PROB and PROF_PROB with the adaptive item selection method SHE_POST perform better while the BOOT method with PWKL or SHE_POST has highest classification rates but longer average test length. For the LH & Len_8 study condition, the same pattern found as for the Med & Len_8. For the LH & Len_12 study condition, the termination rule ATTR_PROB with PWKL or SKE_POST have slightly higher classification accuracy rates than the termination rule PROF_PROB with PWKL or SKE_POST but the latter has shorter average test length. Also, similarly, the BOOT method with PWKL has highest classification accuracy rate but longer test length than others.

For the CSEM_PWKL, it has very similar classification accuracy across the three different adaptive item selections, while the combination with SHE_MARG always has longer average test length than other two across the four study conditions. For the BOOT termination rule, it has highest classification accuracy when the minimum test length is 12 (Len_12) but longer test length. The BOOT termination rule seems to not work well with the SHE_MARG adaptive item selection methods, especially for the shorter minimum test length (Len_8).
Figure 3.

*The Classification Accuracy and Test Length Results for the Mastery-rate Sample.*
**Conditional Sample**

Figure 4 shows the conditional sample results. Note that for the condition sample, the study conditions are only Len_8 and Len_12 because the distribution of profiles/classes are uniform across all possible patterns. The whole figure clearly shows that the ATTR_PROB and the PROF_PROB have very similar patterns in classification accuracy and average test lengths across different item selection methods and study conditions. A general pattern found for the classification accuracy is that the higher profiles (i.e., more attributes are mastered) have higher classification rates and shorter average test length. This observation is not surprising because the nature of the DINA model is not ideal to distinguish low profiles. The DINA model assumes that in order to have a high chance to answer the items correct that the respondent needs to master all attributes measured by the item; therefore, for an incorrect response, the DINA model does not inform which attribute(s) are not mastered.
Figure 4.

*The Classification Accuracy and Test Length Results for the Conditional Sample.*
Summary and Discussion

This study intends to help practitioners better understand the benefit of using the adaptive algorithm for diagnostic assessment and inspire them to use a similar fashion to optimize their algorithm for their own diagnostic tests.

Summary

The adaptive item selection methods investigated in this study, including SHE_POST, SHE_MARG, and PWKL, all perform largely better compared with the RANDOM method—a non-adaptive way of item selection, in terms of better classification accuracy and shorter test length. For a test with a fixed test length 20, the SHE_POST and PWKL have better classification accuracy and similar behavior in terms of item selection. And the same finding were observed for the variable-length adaptive tests given the study conditions investigated.

For the four termination rules investigated in this study, the ATTR_PROB and PROF_PROB generally perform better with either the item selection method PWKL or SHE_POST.

Discussion

The benefit of adaptive diagnostic testing is confirmed in this study when compared with the random method; however, the behavior of adaptive item selection needs more research. One finding regarding this aspect from this study is that, in general, the three adaptive item selection approaches tend to use more simple items for low profiles and more complex items for high
profiles as shown in Figure 1. It indicates that simple items are necessary for low ability students when the items are modeled by DINA.

This is a simulation study and note that the results should not be generalized to the conditions not investigated because the performance of the adaptive algorithm does not only depend on the algorithms themselves, but also to a great extent of the pool quality, the items’ attribute structure, and the model-data fit.

**Future research**

There is still plenty to research in adaptive diagnostic testing. Here, we only list several possible topics that we will conduct for research (but not limited to those). First, we would like to explore the possible effect of different starting points. Second, a stochastic latent attribute model will be used as a data-generation model to generate simulated item response data, which would be more realistic. Last, the BOOT termination rules need to have more research. It should be promising for high-stakes tests that classification is used as part of decision making.
Reference


