the british psychological society
promoting excellence in psychology

British Journal of Mathematical and Statistical Psychology (2020), 73, 88–108

www.wileyonlinelibrary.com

Look-ahead content balancing method in variable-length computerized classification testing

Xiao Li[1]*, Jinming Zhang[1] and Hua-hua Chang[2]

[1]Department of Educational Psychology, University of Illinois at Urbana-Champaign, Illinois, USA

[2]Educational Psychology & Research Methodology, College of Education, Purdue University, West Lafayette, Indiana, USA

Content balancing is one of the most important issues in computerized classification testing. To adapt to variable-length forms, special treatments are needed to successfully control content constraints without knowledge of test length during the test. To this end, we propose the notions of 'look-ahead' and 'step size' to adaptively control content constraints in each item selection step. The step size gives a prediction of the number of items to be selected at the current stage, that is, how far we will look ahead. Two look-ahead content balancing (LA-CB) methods, one with a constant step size and another with an adaptive step size, are proposed as feasible solutions to balancing content areas in variable-length computerized classification testing. The proposed LA-CB methods are compared with conventional item selection methods in variable-length tests and are examined with different classification methods. Simulation results show that, integrated with heuristic item selection methods, the proposed LA-CB methods result in fewer constraint violations and can maintain higher classification accuracy. In addition, the LA-CB method with an adaptive step size outperforms that with a constant step size in content management. Furthermore, the LA-CB methods generate higher test efficiency while using the sequential probability ratio test classification method.

1. Introduction

The computerized classification testing (CCT; Parshall, Spray, Kalohn, & Davey, 2002) method has been applied in a variety of proficiency tests to classify examinees into two or more mutually exclusive groups. Different from the computerized adaptive testing (CAT) method with respect to point estimation of ability, the CCT method does not necessarily acquire an accurate estimate of ability values (Thompson & Prometric, 2007; Weiss & Kingsbury, 1984).

For the purpose of further improving test efficiency, variable-length computerized classification testing (VL-CCT) is adopted (Parshall *et al.*, 2002; Thompson & Prometric, 2007). Variable-length testing refers to tests in which not all examinees receive the same number of items. Before a decision (pass/fail) is made, an examinee with high or low ability who is far from the cut-off score will receive a relatively small number of items compared to an examinee with ability closer to the cut-off score.

The purpose of the VL-CCT method is to provide the decision with as few items as possible, while maintaining decision accuracy at a certain level. The VL-CCT method is a

*Correspondence should be addressed to Xiao Li, Department of Educational Psychology, University of Illinois at Urbana-Champaign, 310 E. Michigan Ave. Apt 10, Urbana, IL 61801, USA (email: xiaoli20@illinois.edu).

DOI:10.1111/bmsp.12165

powerful and efficient approach to classifying examinees into groups using variable test lengths adapted to abilities. It outperforms fixed-length tests in at least three aspects: offering substantially shorter tests than a conventional fixed-length test while maintaining a similar level of classification accuracy (Kingsbury & Weiss, 1983); conforming to the 'equal measurement error variance' with fixed standard error of measurement (SEM) as a stopping rule (Huo, 2009); and allowing subsequent statistical analyses involving measurement errors to be easily handled (Thissen & Mislevy, 2000; Wainer, Dorans, Flaugher, Green, & Mislevy, 2000).

Currently, the VL-CCT method is not as widely adopted as the fixed-length method in educational and psychological assessments for several reasons. First, it is reported that extremely short tests can affect examinees' perceptions of fairness (Huo, 2009; Tonidandel, Quinones, & Adams, 2002). Second, it is difficult to incorporate all statistical and non-statistical constraints into a VL-CCT design. In the VL-CCT implementation, constraints include content balancing, exposure control, answer key balancing, etc. Content balancing refers to the case where a certain proportion of items needs to be selected from each content area. Exposure control means that the item exposure rate should be kept under a specific threshold. Ideally, items should not be over-exposed or under-exposed, in order to protect test security and maximize item pool usage. Answer key balancing means that correct answers should be uniformly distributed among options (Chang & Ying, 1999; Cheng & Chang, 2009; Sympson & Hetter, 1985). However, as the total number of administered items is unknown before a VL-CCT test is terminated, traditional item selection methods cannot accommodate non-statistical constraints properly without pre-specifying a content area range.

The importance of content balancing has been demonstrated by many researchers (Green, Bock, Humphreys, Linn, & Reckase, 1984; Thissen & Mislevy, 2000; Wainer *et al.*, 2000). A number of methods have been proposed to manage non-statistical constraints, including the constraint CAT method (Kingsbury & Weiss, 1983), the modified multinomial model method (Chen & Ankenman, 2004), the modified constraint CAT method (Leung, Chang, & Hau, 2000), the maximum priority index (MPI) (Cheng & Chang, 2009), and the content-weighted item selection index (CWI; Huo, 2009). The CWI method can be adapted to accommodate constraint management in variable-length tests. Furthermore, the MPI method was adjusted and introduced in variable-length multidimensional CAT (Su, 2015, 2016; Yao, 2013). However, it is still a challenging task to control all constraints simultaneously in a variable-length test setting. Thus, it is desirable to develop new content balancing methods that are specifically designed for variable-length tests.

In this paper, we address these challenges by proposing two feasible methods based on a new design, named look-ahead content balancing (LA-CB), which gains control over content coverage in severely constrained VL-CCT programs. Integrated with the MPI item selection method, the two LA-CB based methods simultaneously accommodate non-statistical constraints in VL-CCT. Furthermore, these LA-CB methods are easy to implement in VL-CCT tests. The LA-CB methods are then compared with the MPI and CWI methods with respect to their performance in constraint management and classification accuracy.

The rest of the paper is organized as follows. Four content balancing item selection methods (including two existing and two newly proposed LA-CB methods) and two classification methods are presented in the Section 2. The results of three simulations are summarized in Section 3. Some concluding remarks are made in Section 4, and potential future research directions are discussed in Section 5.

2. Methods

The CCT approach built upon item response theory (IRT) with a three-parameter logistic model (3PLM) (Hambleton & Swaminathan, 1985) is mostly frequently used. The 3PLM defines the probability of an examinee with ability θ answering item j correctly as

$$P_j(X = 1 \mid \theta) = c_j + (1 - c_j)\frac{e^{a_j(\theta - b_j)}}{1 + e^{a_j(\theta - b_j)}}, \tag{1}$$

where a_j is the item discrimination parameter, b_j is the item difficulty parameter and c_j is the guessing parameter or a lower asymptote.

One of the most widely used item selection methods in CCT programs is the maximum Fisher information method (Lord, 1980; Wainer *et al.*, 2000). This selects the next item with the maximum value of Fisher information evaluated at the current ability estimate point $\hat{\theta}$. The Fisher information in the 3PLM is expressed as

$$I_j(\theta) = \frac{(1 - c_j)a_j^2 e^{a_j(\theta - b_j)}}{[1 + e^{a_j(\theta - b_j)}]^2\{1 - c_j + c_j[1 + e^{a_j(\theta - b_j)}]\}}. \tag{2}$$

Other than maximum Fisher information method, the MPI method measures both the information each item carries, and the items' contribution towards meeting constraints.

2.1. Content balancing item selection methods

2.1.1. Maximum priority index

The MPI method is a flexible item selection algorithm that incorporates content balancing constraints in fixed-length CAT. The MPI method heuristically balances constraints in the item selection procedure by including a multiplier in front of an item's Fisher information; and the Fisher information quantifies the contribution to θ estimation. Also, the larger the MPI is, the more desirable it is to administer the item. The priority index of item j is computed as

$$\mathrm{PI}_j = I_j \prod_{k=1}^{K} (\omega_k f_k)^{c_{jk}}, \tag{3}$$

where I_j represents the Fisher information of item j with regard to $\hat{\theta}$. c_{jk} is the indicator for whether constraint k is relevant to item j, and takes the value 1 if k is relevant, and 0 otherwise. ω_k is a predefined weight for constraint k, which is used to quantify the importance of content constraints; that is, major content constraints will receive large weights.

Suppose that the target number for a certain content constraint k is X_k, and that x_k such items have been selected. The resulting scaled quota f_k is

$$f_k = \frac{X_k - x_k}{X_k}. \tag{4}$$

X_k varies over different item selection phases as well as constraints. For example, if content area k involves a lower bound l_k and an upper bound u_k, then X_k equals l_k in the

first phase and u_k in the second. In variable-length tests, the upper limit is bounded by a ratio u_k% and the maximum test length U, which gives $u_k = U \times u_k$%.

To ensure that a sufficient number of items are administered to examinees and thus a reliable test, most VL-CCT programs set both lower and upper bounds on total test length, as well as content area constraints. Therefore, the lower bound of each content area is handled in the first phase and the upper bound in the second phase.

In particular, the desired exposure rate r can be treated as an upper limit in f_k for exposure control purposes, expressed as

$$f_{kr} = \frac{r - n/N}{r}, \tag{5}$$

where r is the exposure rate upper limit, n counts how frequently item j has been administered, and N is the total number of examinees.

2.1.2. Content-weighted item selection index

One content balancing control method proposed for variable-length tests is the CWI method (Huo, 2009). The method incorporates adapted a-stratified methods to control content constraints in variable-length CAT.

Let l_k and u_k denote the lower and upper bounds of the constraint k respectively, and x_k denote the number of selected items from the constraint k. The CWI method is calculated in two phases. In the first phase, the index is expressed as

$$\text{CWI} = \frac{l_k}{l_k - x_k + 1} \mid \hat{\theta} - b \mid. \tag{6}$$

In the second phase, the index is

$$\text{CWI} = \frac{u_k}{l_k - x_k + 1} \mid \hat{\theta} - b \mid. \tag{7}$$

To adjust the CWI method in the variable-length setting with exposure control, the author proposes an adapted a-stratified method. The method selects items from strata in a circularly increasing or decreasing order in the second phase, instead of in a strictly ascending or descending order from the original a-stratified item selection method (Chang & Ying, 1999). This adapted method achieves the best result of all adaptations in Chang and Ying's paper. Therefore, we will continue to use the CWI with the adapted a-stratified method as one of reference methods to compare with the new methods presented below in our simulation studies.

2.1.3. Look-ahead content balancing

The problem with using existing content balancing methods is that the upper bound of each content area is unknown before the test is terminated. In VL-CCT programs, each content balancing constraint usually includes both a lower bound and an upper bound. The lower bounds are fixed values to ensure sufficient items are administered to examinees and the reliability of the test. The upper bounds are usually controlled by a target percentage. As a result, when the total test length is changing, the program cannot

determine the exact upper bounds. The existing MPI method and the CWI method both use maximum upper bounds, which are the total test length times the target percentages, as the target upper bounds in the second phase. However, the maximum upper bounds can be much larger than the actual ones since some tests may terminate early. As a result, the content constraints cannot be controlled properly.

To solve this problem, we first proposed a straightforward solution. The upper bounds are decided by a fixed value called the 'step size'. By looking one step ahead, the upper bounds keep determined by the existing number of selected items plus the step size in each item selection procedure. An alternative method is to determine the step size by a confidence interval (CI) derived from the Fisher information. The upper bounds are then decided in the same way.

We introduced the idea of looking ahead by taking one step forward in both methods. Both of them prove to be reliable in maintaining high test accuracy and content management. In addition, we can use the flexible values of step size to decide the priority of achieving higher classification accuracy or fewer constraint violations. Besides, the Fisher information contributes the measure which further refines the step size's precision in determining upper bounds. The resulting VL-CCT program can show its high test efficiency over fixed-length tests without compromising constraint management.

Specifically, the LA-CB design adopts the idea of a a two-phase item selection strategy (Cheng & Chang, 2009; Cheng, Chang, & Yi, 2007). It handles lower bounds in the first phase and upper bounds in the second. Using the same notation as above, let x_k denote the number of selected items from content area k. The following equations must be satisfied:

$$x_k \geq l_k, \tag{8}$$

and

$$x_k \leq \mathrm{TL} \times u_k\%, \tag{9}$$

where l_k is the lower bound for content area k, $u_k\%$ is the target percentage of content area k in the second phase and TL is total test length. The priority index PI_j is then computed by (3).

In the first phase, we have

$$f_k = \frac{l_k - x_k}{l_k}. \tag{10}$$

So f_k gives the quota of the distance between the lower bound and the current selection length.

In the second phase, because both the total test length and the total number of items received by examinees are changing, we should have a solution to determine what would be the remaining length. The way to go about this is to take one step ahead, by introducing either a constant value or an adaptive value determined by the CI. We call the value the step size, S. Suppose the maximum test length is U, which is larger than or equal to the actual test length TL. Then the target percentage $u_k\%$ must satisfy

$$x_k + S \times u_k\% \leq U \times u_k\%, \tag{11}$$

which gives

$$1 \leq S \leq U - \sum_{k=1}^{K} x_k. \tag{12}$$

Inequalities (11) and (12) indicate that the number of selected items plus S cannot exceed maximum test length in VL-CCT programs. Besides, if the test is still in progress (i.e., the maximum test length is not reached and the termination criterion has not satisfied), at least one item should be selected, in which case the value of S is at least 1. Therefore, the step size S can be a constant integer within the range given by (12). We refer to the LA-CB method with constant step size S^{constant} as LA-CB-C.

To further improve the precision in determining the upper bound, we used the ability confidence interval (ACI) method to predict the step size S. By evaluating the distance between the current Fisher information and desired Fisher information, the value of S is calculated. As a result, constraints under each content area can better controlled. The method is referred to here as LA-CB-A.

By the ACI method, a CI, based on $\hat{\theta}$ and the conditional standard error of measurement SEM, will be constructed and compared to the cut-off score. A CI is expressed as

$$\hat{\theta} - Z_\alpha \times \text{SEM} < \theta < \hat{\theta} + Z_\alpha \times \text{SEM}, \tag{13}$$

where Z_α is the normal deviate for a $100(1 - \alpha)\%$ CI.

To estimate SEM, by the central limit theorem, under local independence and large n assumptions, we have

$$SD(\hat{\theta}) \rightarrow \frac{1}{\sqrt{\sum_{j=1}^{n} I_j(\theta)}}, \quad \text{as} \quad n \rightarrow \infty, \tag{14}$$

where $I_j(\theta)$ represents the Fisher information of item j. After the first k items have been administered, the CI is approximated b

$$\hat{\theta} - Z_\alpha \frac{1}{\sqrt{\sum_{j=1}^{k} I_j(\theta)}} < \theta < \hat{\theta} + Z_\alpha \frac{1}{\sqrt{\sum_{j=1}^{k} I_j(\theta)}}. \tag{15}$$

Since the LA-CB-A method is applied in the second phase, at least l_k items have already been administered. The number of items is large enough that the accumulated Fisher information can be used to approximate SEM.

Denote the cut-off score by θ_0. If the lower bound of the CI equals θ_0, the entire CI will lie to the right of θ_0, leading to the classification of passing the test under the ACI method, where

$$\theta_0 = \hat{\theta} - Z_\varepsilon \frac{1}{\sqrt{\text{FI}_0}}. \tag{16}$$

In contrast, if the upper bound of the CI equals θ_0, the entire CI will lie to the left of θ_0, which gives

$$\theta_0 = \hat{\theta} + Z_\varepsilon \frac{1}{\sqrt{\mathrm{FI}_0}}. \tag{17}$$

The test will terminate and the examinee will be classified as failing the test.

Both equations (16) and (17) give the same total desired Fisher information FI_0 to the terminated test which is calculated as

$$\mathrm{FI}_0 = \left[\frac{Z_\varepsilon}{\theta_0 - \hat{\theta}}\right]^2. \tag{18}$$

Therefore, the number of items to be selected in the next steps has the range

$$\frac{\mathrm{FI}_0 - \sum_{j=1}^{k} I_j(\hat{\theta})}{\max(I_{\text{unselected}})} < \text{no. of items} < \frac{\mathrm{FI}_0 - \sum_{j=1}^{k} I_j(\hat{\theta})}{\min(I_{\text{unselected}})}, \tag{19}$$

where $\max(I_{\text{unselected}})$ and $\min(I_{\text{unselected}})$ represent the maximum and minimum Fisher information based on current $\hat{\theta}$, respectively, for an item in the remaining item pool.

To conservatively control content constraints, the predicted number of remaining items should be as small as possible. Therefore, the LA-CB-A method uses the left bound in (19) as the look-ahead upper bound. The adaptive step size will be calculated as

$$S_0^{\text{adaptive}} = \frac{\mathrm{FI}_0 - \sum_{j=1}^{k} I_j(\hat{\theta})}{\max(I_{\text{unselected}})} - \sum_{k=1}^{K} x_k. \tag{20}$$

When the test is in a relatively early stage, the standard error of estimated ability is large and the accumulated Fisher information is not yet close to FI_0. As a result, the adaptive step size S_0^{adaptive} can be very large. To take advantage of S_0^{adaptive} while having it controlled in a reasonable range, we integrated S_0^{adaptive} with the constant step size S^{constant}. The resulting S^{adaptive} in the LA-CB-A method is calculated by

$$S = \min\{S_0^{\text{adaptive}}, S^{\text{constant}}\}. \tag{21}$$

With the step size S for either the LA-CB-C or LA-CB-A method, the quota f_k is calculated by

$$f_k = \frac{S \times u_k\%}{x_k + S \times u_k\%}. \tag{22}$$

The priority index is calculated by (3) for each item j and the item with the MPI is selected and administered.

FI_0, S, and f_k are iteratively predicted and updated and items are administered following the same procedure until the termination criterion is satisfied or the maximum test length is reached. Examinees are classified as pass/fail based on the classification criterion if the test terminates before the maximum test length is reached. Otherwise, examinees are classified based on the comparison between the estimated ability $\hat{\theta}$ and the cut-off score θ_0.

2.2. Classification methods

2.2.1. Sequential probability ratio test

The sequential probability ratio test (SPRT) (Eggen, 1999; Wald, 1947) turns out to be to be a reliable method in the adaptive test for classifying examinees into categories (Eggen & Straetmans, 2000; Spray & Reckase, 1996). It compares the ratio of the likelihoods of two competing hypotheses. In CCT programs, the likelihood is calculated with the probability of an examinee's response to item i, given the true hypothesis. The probability is calculated with an IRT item response function.

For the purposes of this approach, the statistical hypotheses are formulated as

$$H_0 : \theta \leq \theta_0 - \delta = \theta_1 \tag{23}$$

against

$$H_1 : \theta \geq \theta_0 + \delta = \theta_2, \tag{24}$$

where δ is the indifference zone, accounting for the uncertainty of decisions due to measurement error. The value θ is close to the true ability measure θ_0.

Acceptable decision error rates are then specified as

$$P(\text{accept } H_0 \mid H_0 \text{ is true}) \geq 1 - \alpha, \tag{25}$$

and

$$P(\text{accept } H_0 \mid H_1 \text{ is true}) \leq \beta, \tag{26}$$

where α and β are the nominal Type I and Type II error rates, respectively.

Tests meeting these decision error rates are then implemented using the SPRT. The test statistic used is the ratio between the values of the likelihood functions under the alternative hypothesis and the null hypothesis,

$$\mathrm{LR}(\theta_2, \theta_1; \mathbf{y}) = \frac{L(\theta_2; \mathbf{y})}{L(\theta_1; \mathbf{y})} = \frac{\prod_{j=1}^{K} P_j(\theta_2)^{y_j}[1 - P_j(\theta_2)]^{1-y_j}}{\prod_{j=1}^{K} P_j(\theta_1)^{y_j}[1 - P_j(\theta_1)]^{1-y_j}}, \tag{27}$$

where $\mathbf{y}$ denotes responses $y_1, y_2, \ldots, y_K$ and K denotes the total number of items. $P_j(\theta)$ is the item response function of the 3PLM from equation (1). Large values of this ratio indicate that the examinee's θ is above θ_0, and small values indicate that θ is below θ_0. That is, a statistical test satisfies acceptable decision error rates if it uses the following procedure (Eggen, 1999): if

$$\frac{\beta}{1 - \alpha} < \mathrm{LR}(\theta_2, \theta_1; \mathbf{y}) < \frac{1 - \beta}{\alpha}, \tag{28}$$

the sampling procedure continues; if

$$\mathrm{LR}_k(\theta_2, \theta_1; \mathbf{y}) \leq \frac{\beta}{1 - \alpha}, \tag{29}$$

we accept H_0 and classify the examinee as failing in the test; if

$$\mathrm{LR}_k(\theta_2, \theta_1; \mathbf{y}) \geq \frac{1-\beta}{\alpha}, \tag{30}$$

we reject H_0 and classify the examinee as passing the test.

2.2.2. Ability confidence interval

The ACI method is an alternative way to make a classification decision. A 95% CI is constructed around the examinee's estimated theta after each item administered. If the examinee's 95% CI is above the cut-off score θ_0, then the examinee passes the test. If the CI falls below θ_0, then the examinee fails. If θ_0 is equal to or within the examinee's CI, then the test will continue until a pass/fail decision can be made or the maximum test length is reached.

3. Simulation studies and results

3.1. Overview

Three simulation studies were conducted for this paper. In the first simulation study, the evaluation between the ACI and SPRT methods is based on classification accuracy and test efficiency criteria in the application of the LA-CB-C method. The main purpose of the first study is to choose a preferable classification method in the current setting so that the preferred one would be applied in the following two simulation studies. Only the results of LA-CB-C method are presented in paper article since the LA-CB-A method produces similar results. In the second simulation study, we evaluate whether the LA-CB-C method controls content constraints better than the existing MPI and CWI methods in VL-CCT tests, where baselines are taken to be the MFI without exposure control and the randomized method. The comparisons are conducted with respect to multiple perspectives, including classification accuracy, test efficiency, content balancing and exposure control. In the third simulation study, we examine whether the LA-CB-A method further improves the content balancing performance on top of the LA-CB-C method. Details of the settings and the results of the three studies are discussed in the following subsections.

3.2. Data generation

3.2.1. Item pool structure

A hypothetical item bank is simulated under the 3PLM with 400 items, partitioned into four stages with parameter a evenly distributed at 0.5, 1.0, 1.5, 2.0. Other item parameters are generated as $b \sim N[0, 1]$ and $c \sim U[0, 0.25]$. The item bank is evenly divided into four content areas, each of which contains 100 items. Each content area is assumed with 25% desired selection rate. The four content areas are considered equally important and the weights are all set to 10. The minimum and maximum test lengths are set at 28 and 60. Therefore, for each content area, the number of selected items under each constraint k ($k = 1, 2, 3, 4$) should be bounded between integers 7 and 15.

As for test security purposes, the exposure rate of all items is required to be controlled under 0.2, which means items are administered to no more than 20% of examinees. The constraint is expressed in equation (5). Because the simulated test is considered high-stakes, the weight of the exposure control constraint is set to 100.

3.2.2. Examinee generation

We drew 2,000 θs from $N[0, 1]$ as our simulated examinees. In order to mitigate the randomness in the results, 20 replications were performed for each of the 18 step sizes of the LA-CB-C and LA-CB-A methods, and for each of the other four item selection methods, using the same item bank and generated examinees in the second and third simulation studies. The averaged results were presented. The pass rate of the test is taken to be 50%.

3.2.3. Model settings

The indifference region δ for SPRT method is set to 0.2. The cut-off score for θ_0 is 0. As a result, θ_1 and θ_2 are -0.2 and 0.2, respectively. Parameters α and β for SPRT are set to 0.05. In addition, 18 integral values are generated for the step size S in the LA-CB methods, ranging from 3 to 20. As S should be bounded in the range given by (12) in both LA-CB-C and LA-CB-A methods, in the later stage of the test, the constant step size we generate, S^{constant}, may exceed $U - \sum_{k=1}^{K} x_k$ as more and more items are selected. The actual step size S^{actual} that we use in the study is calculated by S from the LA-CB-C and LA-CB-A as

$$S^{\text{actual}} = \max\left\{1, \min\left\{S, U - \sum_{k=1}^{K} x_k\right\}\right\}, \tag{31}$$

where $S = S^{\text{constant}}$ in LA-CB-C and $S = S^{\text{adaptive}}$ in LA-CB-A.

At the beginning of the test, the first three items are always selected randomly because we lack the knowledge to compute the Fisher information. The following items are selected from the two best items, where the best item refers to the one with maximized priority index, Fisher information, or minimized weighted index, depending on which method is used.

3.3. Evaluation criteria

Various criteria are used to analyze and compare the two newly proposed methods with traditional methods. Results are evaluated based on the following four main aspects. Note that the last criterion is for the first simulation study only.

1. *Classification accuracy*. Three criteria are used for classification accuracy comparison in the simulations: classification error rate (CER), Type I error rate (Type I ER) and Type II error rate (Type II ER). Meanwhile, mean square error is also calculated as a measurement precision criterion, given by

$$\text{MSE} = \frac{\sum_{i=1}^{N}(\hat{\theta}_i - \theta_i)^2}{N} \tag{32}$$

2. *Content balancing*. The total numbers of violated content constraints across various examinees are evaluated as a criterion for content balance. Denote by V_i the total number of constraints violated in all content areas for examinee i. The average number of constraints violated in a test is calculated by

$$\bar{V} = \frac{\sum_{i=1}^{N} V_i}{N}, \tag{33}$$

where N denotes the total number of examinees (which equals 2,000). The average value of $\bar{V}$ across different examinees, the maximum $\bar{V}$ and minimum $\bar{V}$ are given for comparison. The grand average $\bar{V}$ is defined as

$$\text{Average } \bar{V} = \frac{\sum_{p=P_0}^{P} \bar{V}_p}{P - P_0 + 1}, \tag{34}$$

where $\bar{V}_p$ denotes the average number of constraints violated for the pth step size and P_0 and P denote respectively the minimum and maximum step sizes we generated for LA-CB-C and LA-CB-A models. Maximum and minimum $\bar{V}$ are also calculated across different step sizes.

3. *Exposure control*. Four criteria are used for the purpose of evaluating exposure control across five different methods. They are the maximum item exposure rate, the proportion of over-exposed items (items with exposure rate higher than 0.2), the proportion of unused items, and χ^2. χ^2 is designated to measure the similarity between observed and expected exposure rates (ER),

$$\chi^2 = \sum_{j=1}^{K} \frac{(\text{ER}_j - \overline{\text{ER}})^2}{\overline{\text{ER}}}, \tag{35}$$

where j denotes the jth item, K denotes the total number of items, and $\overline{\text{ER}}$ shows the average exposure rate of all the items in the pool.

4. *Test efficiency*. To compare the test efficiency between two classification methods (ACI and SPRT) in the first study, the average test lengths TL across various examinees are calculated, conditioning on different step sizes. $\overline{\text{TL}}$ is expressed as:

$$\overline{\text{TL}} = \frac{\sum_{i=1}^{N} \text{TL}_i}{N}, \tag{36}$$

where TL_i denotes the test length for the ith examinee and N = 2,000 is the total number of examinees.

3.4. Results of simulation 1

The ACI and SPRT classification methods are adopted with the LA-CB-C method and compared for classification accuracy and test efficiency. The classification accuracy includes the CER, Type I ER and Type II ER. Therefore both the classification specificity and sensitivity can be shown. A reliable classification method is expected to provide both low classification error rate and short average test length.

The focus of the first study is to find out the most appropriate classification method, which is a critical part of the VL-CCT design so that LA-CB methods can be further investigated on top of the recommended method. The classification method with the better performance is applied in the following two studies.

Table 1 gives a comparison of classification accuracy and test efficiency between the SPRT and ACI methods. The average test lenth shows their performance in improving test efficiency with the benefit of variable length setting. While the average test length with the SPRT method is 29.8, that with the ACI method is 37.1. With SPRT as the termination criterion, the average test length is 19.6% shorter than with ACI. Figure 1 compares total test lengths between the two methods conditional on 18 step sizes. In addition, 18 one-way ANOVA tests were run to compare the test lengths generated by the ACI and SPRT methods conditional on 18 step sizes. All p-values were reported to be less tgan than $2 \times e^{-16}$, indicating that the test lengths generated by the ACI and SPRT methods are significantly different.

The second row in Table 1 gives the average number of constraints violated in tests ($\bar{V}$). The value is 0.011 with SPRT and 0.017 with ACI. The result accords with the result given for test length in Table 1 and Figure 1, since the ACI method tends to give a longer test so there is a higher probability of constraint violation.

The last part of the table presents the overall CERs of the two methods. The ACI method gives a slightly better performance with 6.0% error rate, while SPRT has 6.3% error rate on classifying examinees. The difference of the average CERs between the ACI and SPRT methods is thus only 0.3%. Figure 2 presents the CERs of the two methods conditional on 18 step sizes, while Figures 3 and 4 give corresponding Type I and Type II ERs. The fluctuations of the curves are due to randomness from the test setting. Items are selected randomly from the two best ones, and ability estimation errors also result in randomness in item selection procedure. In general, the differences in CERs between the

Table 1. Overall performance of sequential probability ratio test (SPRT) and ability confidence interval (ACI) classification methods

Methods	SPRT	ACI
Avg. test length TL	29.85	37.11
Grand avg. violated constraints $\bar{V}$	0.011	0.017
Average classification error rate	0.063	0.060
Average Type I error rate	0.033	0.030
Average Type II error rate	0.030	0.030

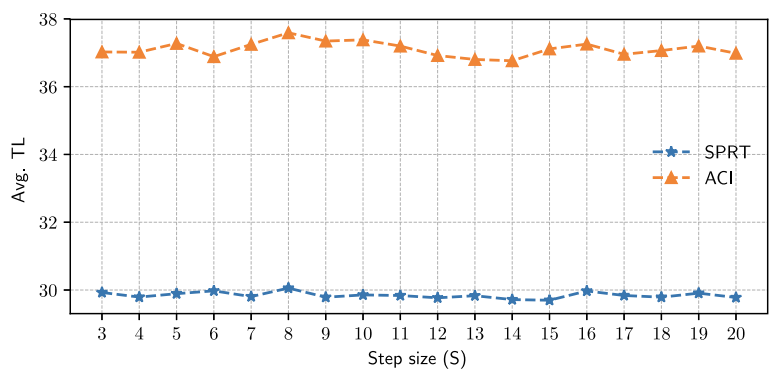

Figure 1. Average test length (TL) of sequential probability ratio test (SPRT) and ability confidence interval (ACI). [Colour figure can be viewed at wileyonlinelibrary.com]

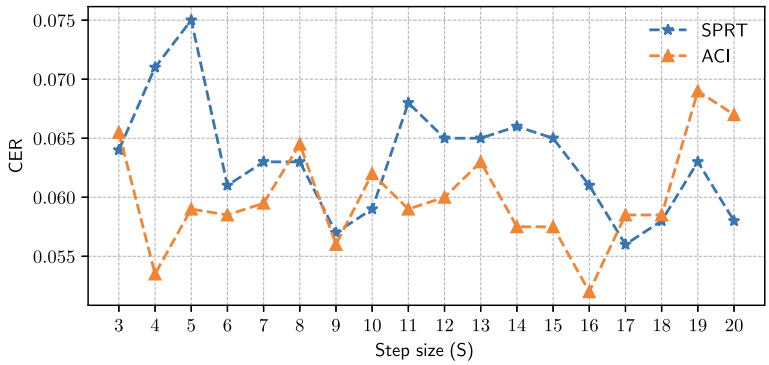

Figure 2. Classification error rate (CER) of sequential probability ratio test (SPRT) and ability confidence interval (ACI). [Colour figure can be viewed at wileyonlinelibrary.com]

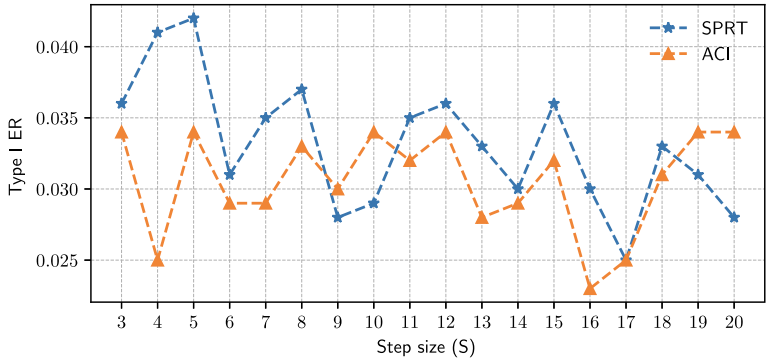

Figure 3. Type I error rate (ER) of sequential probability ratio test (SPRT) and ability confidence interval (ACI). [Colour figure can be viewed at wileyonlinelibrary.com]

two methods across 18 step sizes are quite small. The results show that similar classification accuracy is achieved by the two methods.

The results given in Table 1 and Figures 1–4 clearly show that SPRT improves test efficiency by shortening the test length by 19.6% without losing much capacity to maintain high classification accuracy, which is 93.7% here. A similar conclusion that SPRT tends to gives a better performance in CCT can be found in other work (Babcock & Weiss, 2009; Lin, 2011; Thompson, 2009) as well. Thus, SPRT is shown to be an efficient classification method which is used in simulations 2 and 3 for a further evaluation of the LA-CB methods.

3.5. Results of simulation 2

Eighteen step sizes are generated for a comparison of the LA-CB-C method with the CWI, MPI and the baseline of MFI and randomized item selection method. The influence of different step sizes will be evaluated from different perspectives mentioned above. The SPRT is adopted here as the classification method.

Since the LA-CB-C method is designed as a content balancing item selection method without sacrificing classification accuracy, criteria including classification accuracy,

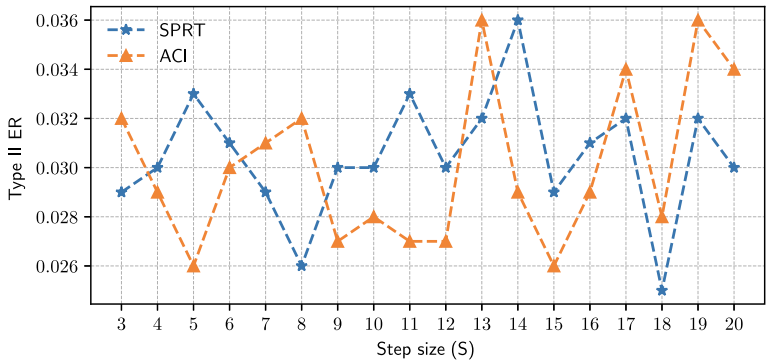

Figure 4. Type II error rate (ER) of sequential probability ratio test (SPRT) and ability confidence interval (ACI). [Colour figure can be viewed at wileyonlinelibrary.com]

content balancing and exposure control are recorded for comparison. We replicated the simulation 20 times and averaged the resulting values to give a reliable result.

Figures 5 and 6 present the CERs and the numbers of constraints violated across different step sizes with the LA-CB-C method. The trendline is a linear regression line which gives the linear trend of those two. Obviously, as the step size increases, the classification accuracy rate improves slightly, with error rate decreasing (see Figure 5). At the same time, the number of violated constraints increases greatly with larger step sizes (see Figure 6). There clearly exists a trade-off between classification accuracy and content balancing regarding different step sizes. With a decreasing step size, content constraints of selected items can be better controlled, with a slight loss of classification accuracy.

Table 2 and Figure 7 present the classification accuracy achieved by the LA-CB-C method for 18 step sizes and the MPI and CWI methods, compared to the baseline of MFI and randomized item selection methods. The results show that the randomized method has the highest CER, the MFI method has the lowest, while the CERs of the LA-CB-C, MPI and CWI methods lie in between. The average CER of the LA-CB-C method is 6.4%, slightly higher than the MPI method's 6.2% but much lower than the CWI method's 7.8%. There is no obvious optimum step size that achieves the lowest error rate. CERs of the LA-CB-C method under different step sizes all lie within the range 6.2–6.6%.

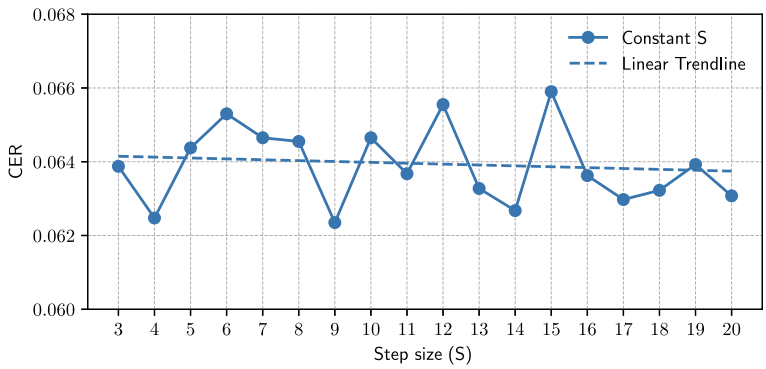

Figure 5. The LA-CB-C method classification error rate (CER) with linear trendline. [Colour figure can be viewed at wileyonlinelibrary.com]

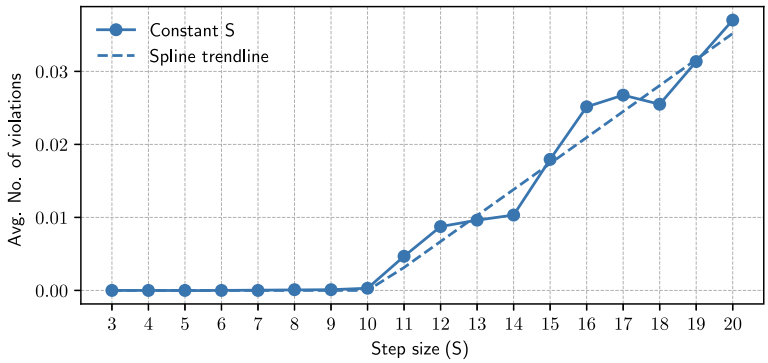

Figure 6. Average number of constraint violations ($\bar{V}$) of the LA-CB-C method with spline trendline. [Colour figure can be viewed at wileyonlinelibrary.com]

Meanwhile, Figure 6 shows that the average number of constraints violated ($\bar{V}$) in the LA-CB-C method with different step sizes is always under 0.040, while $\bar{V}$ for the MPI method is 0.054, the MFI method is 8.14, the CWI method is 2.23 and the randomized method is 7.02 (see Table 5). In particular, with step size from 3 to 10, the LA-CB-C method has no constraint violation. Even with step size 20, which has the highest $\bar{V}$, the LA-CB-C method still performs better on constraint management than the other four

Table 2. Classification error rates (ER) and mean square error (MSE) of the LA-CB-C method and three other methods for 18 step sizes *S*

S	CER	Type I ER	Type II ER	MSE
3	0.064	0.034	0.030	0.074
4	0.062	0.033	0.030	0.074
5	0.064	0.033	0.031	0.075
6	0.065	0.034	0.031	0.075
7	0.065	0.033	0.031	0.075
8	0.065	0.033	0.032	0.074
9	0.062	0.033	0.030	0.075
10	0.065	0.034	0.031	0.076
11	0.064	0.032	0.032	0.075
12	0.066	0.034	0.031	0.075
13	0.063	0.033	0.030	0.075
14	0.063	0.033	0.030	0.075
15	0.066	0.034	0.032	0.075
16	0.064	0.033	0.031	0.074
17	0.063	0.033	0.030	0.075
18	0.063	0.033	0.030	0.075
19	0.064	0.032	0.032	0.075
20	0.063	0.033	0.030	0.075
LA-CB-C average	0.064	0.033	0.031	0.075
Maximum priority	0.062	0.032	0.030	0.074
Content-weighted	0.078	0.040	0.039	0.118
Maximum information	0.054	0.032	0.022	0.054
Randomized	0.084	0.042	0.042	0.254

methods. This shows that the LA-CB-C method significantly improves content balancing compared to the MPI and CWI methods.

Table 3 shows that the exposure rate is well controlled by both the LA-CB-C and MPI methods, especially compared with the MFI method. The maximum exposure rates of the LA-CB-C and MPI methods are both under 0.2, and all items in the pool are used.

The results show that the LA-CB-C method has classification accuracy comparable to the MPI and MFI methods, better than the CWI method. In addition, the LA-CB-C method generates far fewer violated constraints than the other four methods, while exhibiting similar exposure control performance with the MPI method but better than the CWI, MFI and randomized methods. The results indicate that from the content constraint management perspective, the LA-CB-C method outperforms all the other methods.

Table 3. Overall exposure control indices

Methods	LA-CB-C	Maximum priority	Content-weighted	Maximum information	Randomized
Max. exposure rate	0.178	0.175	0.166	0.532	0.100
Over-exposed (%)	0	0	0	3.2	0
Never exposed (%)	0	0	0	0	0
χ^2	20.297	20.228	4.009	83.041	0.153

Table 4. LA-CB-A classification error rates (ER) and mean square error (MSE) for different step sizes *S*

Constant *S*	CER	Type I ER	Type II ER	MSE
3	0.065	0.034	0.031	0.076
4	0.063	0.032	0.031	0.076
5	0.064	0.034	0.030	0.076
6	0.064	0.033	0.031	0.075
7	0.062	0.032	0.030	0.075
8	0.064	0.033	0.031	0.074
9	0.065	0.033	0.031	0.076
10	0.063	0.033	0.031	0.073
11	0.065	0.033	0.031	0.074
12	0.062	0.032	0.030	0.074
13	0.066	0.034	0.032	0.075
14	0.063	0.034	0.029	0.074
15	0.063	0.032	0.031	0.076
16	0.064	0.033	0.031	0.074
17	0.064	0.033	0.031	0.074
18	0.062	0.032	0.030	0.075
19	0.064	0.033	0.030	0.074
20	0.064	0.033	0.031	0.075
LA-CB-A average	0.064	0.033	0.031	0.075
Maximum priority	0.062	0.032	0.030	0.074
Content-weighted	0.078	0.040	0.039	0.118
Maximum information	0.054	0.032	0.022	0.054
Randomized	0.084	0.042	0.042	0.254

3.6. Results of simulation 3

The LA-CB-A method is designed to improve the LA-CB-C method in terms of meeting the content constraints. The criteria to compare the LA-CB-A and LA-CB-C methods include classification accuracy, content balancing, and exposure control. The simulation was replicated 20 times and the results are summarized in Tables 4 and 5 and Figures 7 and 8.

The adaptive step size is used in the LA-CB-A method, expected to better control the content area constraints based on the LA-CB-C method. Table 4 and Figure 7 show the classification accuracy of the LA-CB-A method compared to other methods. The average CER of the LA-CB-A is 6.4%, the same as that of the LA-CB-C method, close to the MPI method, slightly higher than the MFI method, and much lower than the CWI and randomized methods.

Table 5 and Figure 8 present the overall content balancing performance achieved by the LA-CB-C and LA-CB-A methods, compared to the other four methods. Obviously, the LA-CB-C and LA-CB-A methods both outperform other methods, while the LA-CB-A method has the smallest average $\bar{V}$ (see Table 5). Looking in detail at the two methods for different step sizes, the LA-CB-A method controls $\bar{V}$ better than the LA-CB-C method (with smaller $\bar{V}$), especially when the step size gets larger. This makes sense since more constraints tend to be violated when the step size gets larger, while the LA-CB-A method gives a look-ahead prediction of the test length with an adaptive step size, which is no larger than the constant step size in LA-CB-C. It is also worth noting that both the LA-CB-C and LA-CB-A methods control constraints almost perfectly when the step size is smaller

Table 5. Summary of content constraint violations ($\bar{V}$)

Measures	Average $\bar{V}$	Max $\bar{V}$	Min $\bar{V}$
LA-CB-C	0.0110	0.0370	0
LA-CB-A	0.0102	0.0319	0
Maximum priority*	0.0540	–	–
Content-weighted*	2.2295	–	–
Maximum information*	8.1380	–	–
Randomized*	7.0230	–	–

Note. This table summarizes the statistics of $\bar{V}$ for 18 step sizes. Methods with (*) do not include step sizes to make item selections and therefore maximum and minimum $\bar{V}$ are not applicable.

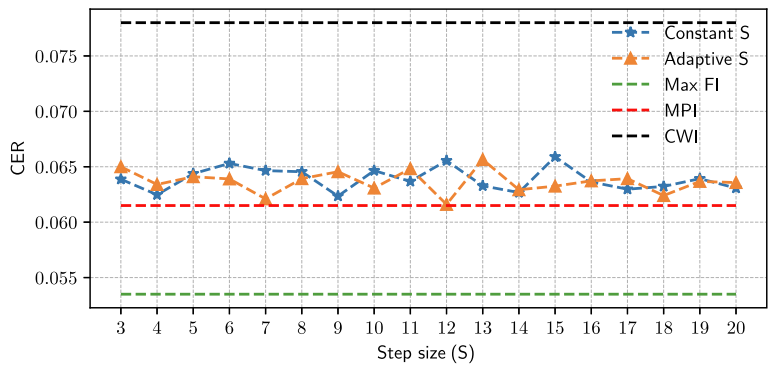

Figure 7. Classification error rate (CER) of the LA-CB methods with constant (Constant *S*) and adaptive (Adaptive *S*) step sizes and maximum priority (MPI), maximum Fisher information (Max FI) and content-weighted (CWI) methods. [Colour figure can be viewed at wileyonlinelibrary.com]

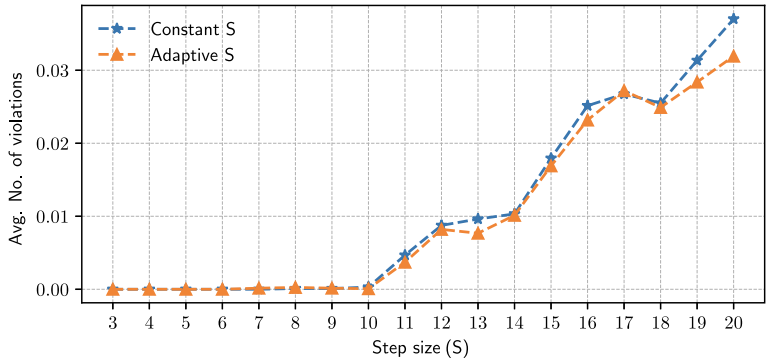

Figure 8. Average number of content constraint violations ($\bar{V}$) of the LA-CB methods with constant (Constant *S*) and adaptive (Adaptive *S*) step sizes. [Colour figure can be viewed at wileyonline library.com]

Table 6. Overall exposure control indices

Methods	LA-CB-C	LA-CB-A
Maximum exposure rate	0.178	0.178
Over-exposed (%)	0	0
Never exposed (%)	0	0
χ^2	20.297	20.289

than 11 (Figure 8). Table 6 shows that the LA-CB-A method also controls the exposure rate very well and is comparable to the LA-CB-C method.

The results show that the LA-CB-A method does manage constraints better than the LA-CB-C method with high classification accuracy and low exposure rate. The LA-CB-A method improves the control of content constraints significantly, especially for larger step sizes, and gives perfect constraint management for small step sizes.

4. Conclusions

The results reported in the preceding section indicate that the proposed LA-CB methods with SPRT are promising solutions to content constrained VL-CCT tests. First, the results show that the LA-CB methods perform better than the CWI and MPI methods in controlling constraints (e.g., content area constraints and exposure rate), while still maintaining high classification accuracy. Second, with adaptive step sizes, the trade-off between the classification accuracy and constraint management can be alleviated. Specifically, the LA-CB methods reduce the number of constraint violations without sacrificing classification accuracy. Third, both the LA-CB-C and LA-CB-A methods are flexible and easy to implement in practice.

The VL-CCT program shows its advantages in improving test efficiency and accuracy. Yet, due to the lack of information on test length, the non-statistical constraints are hard to control, which is very different from the fixed-length CCT program. Now with the proposed LA-CB methods, it is possible to control content constraints and achieve high classification accuracy simultaneously. As such, the VL-CCT approach can play a more important role in future large-scale tests.

5. Future directions

There are several interesting research directions that are worth exploring in the future. Firstly, different stopping rules can be evaluated and optimally determined, and the LA-CB methods can be adjusted based on the preferred stopping rule (Babcock & Weiss, 2009). It is worth examining whether the LA-CB methods can be further improved with other termination criteria including the SPRT stopping rule (van Groen, Eggen, & Veldkamp, 2016), the generalized likelihood ratio (Thompson, 2011), the fixed SEM stopping rule (Choi, Grady, & Dodd, 2011), the information stopping rule (Chang & Ying, 2004), and the projection-based stopping rules (Luo, Kim, & Dickison, 2018).

Secondly, a variation of the LA-CB method integrated with the shadow test approach (van der Linden, 1998) shall be investigated. Different methods dealing with content constraints are proposed and examined using the shadow test approach in fixed-length CAT (van der Linden, 2009; van der Linden & Chang, 2003). To this end, it could be helpful to examine a modified LA-CB method combined with the shadow test approach in a variable-length test setting and compare it with the LA-CB methods proposed in this paper.

Thirdly, it would also be interesting to examine how the LA-CB methods work when integrated with other item selection methods including the variable-length modified multinomial model method (Chen & Ankenman, 2004), the content-weighted item selection index method (Huo, 2009), and a-stratified method (Chang, Qian, & Ying, 2001; Chang & Ying, 1999). The LA-CB methods could also be extended to multidimensional CAT and examined with existing constraint management methods (Born & Frey, 2017). The adjusted LA-CB methods could be analyzed and compared with other traditional constrained item selection methods from the perspectives of measurement precision, constraint management and exposure control.

References

Babcock, B., & Weiss, D. J. (2009). Termination criteria in computerized adaptive tests: Variable-length CATS are not biased. In *Proceedings of the 2009 GMAC conference on computerized adaptive testing* (vol. *14*). Retrieved from http://iacat.org/sites/default/files/biblio/cat09babcock.pdf

Born, S., & Frey, A. (2017). Heuristic constraint management methods in multidimensional adaptive testing. *Educational and Psychological Measurement*, *77*, 241–262. https://doi.org/10.1177/0013164416643744

Chang, H.-H., Qian, J., & Ying, Z. (2001). A-stratified multistage computerized adaptive testing with b blocking. *Applied Psychological Measurement*, *25*, 333–341. https://doi.org/10.1177/01466210122032181

Chang, H.-H., & Ying, Z. (1999). A-stratified multistage computerized adaptive testing. *Applied Psychological Measurement*, *23*, 211–222. https://doi.org/10.1177/01466219922031338

Chang, Y.-c. I., & Ying, Z. (2004). Sequential estimation in variable length computerized adaptive testing. *Journal of Statistical Planning and Inference*, *121*, 249–264. https://doi.org/10.1016/S0378-3758(03)00119-8

Chen, S.-Y., & Ankenman, R. D. (2004). Effects of practical constraints on item selection rules at the early stages of computerized adaptive testing. *Journal of Educational Measurement*, *41*, 149–174. https://doi.org/10.1111/j.1745-3984.2004.tb01112.x

Cheng, Y., & Chang, H.-H. (2009). The maximum priority index method for severely constrained item selection in computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology*, *62*, 369–383. https://doi.org/10.1348/000711008X304376

Cheng, Y., Chang, H.-H., & Yi, Q. (2007). Two-phase item selection procedure for flexible content balancing in cat. *Applied Psychological Measurement*, *31*, 467–482. https://doi.org/10.1177/0146621606292933

Choi, S. W., Grady, M. W., & Dodd, B. G. (2011). A new stopping rule for computerized adaptive testing. *Educational and Psychological Measurement*, *71*(1), 37–53. https://doi.org/10.1177/0013164410387338

Eggen, T. J. H. M. (1999). Item selection in adaptive testing with the sequential probability ratio test. *Applied Psychological Measurement*, *23*, 249–261. https://doi.org/10.1177/01466219922031365

Eggen, T. J. H. M., & Straetmans, G. J. J. M. (2000). Computerized adaptive testing for classifying examinees into three categories. *Educational and Psychological Measurement*, *60*, 713–734. https://doi.org/10.1177/00131640021970862

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement*, *21*, 347–360. https://doi.org/10.1111/j.1745-3984.1984.tb01039.x

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer-Nijhoff Publishing.

Huo, Y. (2009). *Variable-length computerized adaptive testing: adaptation of the a-stratified strategy in item selection with content balancing*. University of Illinois at Urbana-Champaign. Retrieved from http://hdl.handle.net/2142/14715

Kingsbury, G. G., & Weiss, D. J. (1983). A comparison of IRT-based adaptive mastery testing and a sequential mastery testing procedure. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 257–283). New York, NY: Academic Press.

Leung, C.-K., Chang, H.-H., & Hau, K.-T. (2000). *Content balancing in stratified computerized adaptive testing designs*. ERIC Clearinghouse. Retrieved from https://eric.ed.gov/?id=ED442846

Lin, C.-J. (2011). Item selection criteria with practical constraints for computerized classification testing. *Educational and Psychological Measurement*, *71*(1), 20–36. https://doi.org/10.1177/0013164410387336

Lord, F. (1980). *Application of item response theory to practical testing problems*. hillsdale, nj, lawrence erlbaum ass.

Luo, X., Kim, D., & Dickison, P. (2018). Projection-based stopping rules for computerized adaptive testing in licensure testing. *Applied Psychological Measurement*, *42*, 275–290. https://doi.org/10.1177/0146621617726790

Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. C. (2002). *Practical considerations in computer-based testing*. New York, NY: Springer-Verlag.

Spray, J. A., & Reckase, M. D. (1996). Comparison of sprt and sequential bayes procedures for classifying examinees into two categories using a computerized test. *Journal of Educational and Behavioral Statistics*, *21*, 405–414. https://doi.org/10.3102/10769986021004405

Su, Y.-H. (2015). The performance of the modified multidimensional priority index for item selection in variable-length MCAT. In L. A. van der Ark, D. M. Bolt, W.-C. Wang, J. A. Douglas, & S.-M. Chow (Eds.), *Quantitative psychology research* (pp. 89–97). Switzerland: Springer.

Su, Y.-H. (2016). A comparison of constrained item selection methods in multidimensional computerized adaptive testing. *Applied Psychological Measurement*, *40*, 346–360. https://doi.org/10.1177/0146621616639305

Sympson, J., & Hetter, R. (1985). Controlling item-exposure rates in computerized adaptive testing. In *Proceedings of the 27th annual meeting of the military testing association* (pp. 973–977). https://doi.org/10.2307/1165348

Thissen, D., & Mislevy, R. J. (2000). Testing algorithms. *Computerized Adaptive Testing: A Primer*, *2*, 101–133. https://doi.org/10.4324/9781410605931

Thompson, N. A. (2009). Item selection in computerized classification testing. *Educational and Psychological Measurement*, *69*, 778–793. https://doi.org/10.1177/0013164408324460

Thompson, N. A. (2011). Termination criteria for computerized classification testing. *Practical Assessment, Research & Evaluation*, *16*(4), 2. Retrieved from http://pareonline.net/getvn.asp?v=16&n=4

Thompson, N. A., & Prometric, T. (2007). A practitioners guide for variable-length computerized classification testing. *Practical Assessment Research & Evaluation*, *12*(1), 1–13. Retrieved from http://pareonline.net/getvn.asp?v=12&n=3

Tonidandel, S., Quinones, M. A., & Adams, A. A. (2002). Computer-adaptive testing: The impact of test characteristics on perceived performance and test takers' reactions. *Journal of Applied Psychology*, *87*, 320–332. https://doi.org/10.1037/0021-9010.87.2.320

van Groen, M. M., Eggen, T. J., & Veldkamp, B. P. (2016). Multidimensional computerized adaptive testing for classifying examinees with within-dimensionality. *Applied Psychological Measurement*, *40*, 387–404. https://doi.org/10.1177/0146621616648931

van der Linden, W. J. (1998). Optimal assembly of psychological and educational tests. *Applied Psychological Measurement*, *22*, 195–211. https://doi.org/10.1177/01466216980223001

van der Linden, W. J. (2009). Constrained adaptive testing with shadow tests. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 31–55). New York, NY: Springer.

van der Linden, W. J., & Chang, H.-H. (2003). Implementing content constraints in alphastratified adaptive testing using a shadow test approach. *Applied Psychological Measurement*, *27*, 107–120. https://doi.org/10.1177/0146621602250531

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., & Mislevy, R. J. (2000). *Computerized adaptive testing: A primer*. Abingdon, UK: Routledge. https://doi.org/10.4324/9781410605931

Wald, A. (1947). *Sequential analysis*. Oxford, UK: John Wiley.

Weiss, D. J., & Kingsbury, G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, *21*, 361–375. https://doi.org/10.1111/j.1745-3984.1984.tb01040.x

Yao, L. (2013). Comparing the performance of five multidimensional cat selection procedures with different stopping rules. *Applied Psychological Measurement*, *37*, 3–23. https://doi.org/10.1177/0146621612455687

Received 7 May 2018; revised version received 6 October 2018