Constructing Shadow Tests in Variable-Length Adaptive Testing

Applied Psychological Measurement
2018, Vol. 42(7) 538–552
© The Author(s) 2018
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0146621617753736
journals.sagepub.com/home/apm

Qi Diao¹ and Hao Ren¹

Abstract

Imposing content constraints is very important in most operational computerized adaptive testing (CAT) programs in educational measurement. Shadow test approach to CAT (Shadow CAT) offers an elegant solution to imposing statistical and nonstatistical constraints by projecting future consequences of item selection. The original form of Shadow CAT presumes fixed test lengths. The goal of the current study was to extend Shadow CAT to tests under variable-length termination conditions and evaluate its performance relative to other content balancing approaches. The study demonstrated the feasibility of constructing Shadow CAT with variable test lengths and in operational CAT programs. The results indicated the superiority of the approach compared with other content balancing methods.

Keywords

adaptive testing, termination rule, shadow test approach, variable test length

Computerized adaptive testing (CAT) has been widely used in many disciplines and rigorously studied in the educational measurement arena. One of the important and practical issues in building an operational CAT system in educational measurement is related to implementing content balancing and imposing statistical and nonstatistical constraints on item selection while ensuring the efficiency and measurement precision of the test.

Several content balancing methods have been studied in the literature, such as the weighted deviation method (WDM; Stocking & Swanson, 1993), maximum priority index (MPI; Cheng & Chang, 2009), shadow test approach to CAT (Shadow CAT; van der Linden & Reese, 1998), normalized weighted absolute deviation heuristic (NWADH; Luecht, 1998), and the weighted penalty model (WPM; Shin, Chien, Way, & Swanson, 2009). Among the methods, WDM, MPI, NWADH, and WPM are heuristic. In contrast, Shadow CAT is based on linear programming.

The goal of content balancing is primarily to impose the same test specifications and blueprint requirements for all test takers. However, this task is challenging because optimal ability estimation requires sequential item-level adaptation, whereas constraint realization typically requires simultaneous selection. One solution provided by Shadow CAT is to project future consequences of item selection assuming a fixed-length test. During a CAT administration, a sequence of full-length “shadow tests” are assembled in real time; each shadow test satisfies all

¹ACT, Inc., Monterey, CA, USA

Corresponding Author:

Qi Diao, ACT, Inc., 1 Lower Ragsdale Drive, Suite #150, Monterey, CA 93940, USA.
Email: qi.diao@act.org

statistical and nonstatistical constraints, includes all previously administered items, and provides maximum information at the current ability estimate (van der Linden & Glas, 2010). It has been shown that Shadow CAT outperforms heuristic content balancing methods in terms of providing the optimal solution to CAT (e.g., He, Diao, & Hauser, 2014; Patton, Diao, & Boughton, 2013).

One important component of CAT is termination rules, that is, how to end each examinee's test. Several termination rules have been proposed in the literature. They generally fall into two main categories: fixed-length termination and variable-length termination (Weiss & Kingsbury, 1984). Under a fixed-length termination condition, the CAT procedure terminates when a fixed number of items have been administered. In comparison, variable-length approaches generally aim to end the CAT when as soon as a prespecified level of measurement precision has been achieved, which potentially allows all examinees to achieve the same level of measurement precision. There are several methods for variable-length termination rules in the literature, such as the standard error (*SE*) termination rule (Weiss & Kingsbury, 1984), Minimum Fisher Information (MFI) termination rule (Gialluca & Weiss, 1979; Maurelli & Weiss, 1981), and predictive standard error reduction (PSER) termination rule (Choi, Grady, & Dodd, 2011). Termination rules have also been proposed for computerized classification testing, such as sequential probability ratio testing (SPRT; Reckase, 1983; Wald, 1947) and generalized likelihood ratio (GLR; Thompson, 2009). Variable-length termination may also place an upper/lower limit on the number of items to administer to prevent unexpectedly long/short tests. Compared with fixed-length termination, variable-length termination can provide better measurement efficiency and provide fair tests by ensuring equal measurement precision for all examinees.

Shadow CAT has previously been used with fixed-length tests, although van der Linden (2005) has mentioned the possibility of using Shadow CAT for variable-length tests. However, no details were provided as to how to construct such tests apart from the traditional Shadow CAT framework that assumes a fixed-length test. Under the original conceptualization of Shadow CAT, a full-length test is constructed for each item selection, satisfying all constraints and requirements (called a shadow test), using an automated test assembly (ATA) model (van der Linden, 2005). The primary goal of this study is to develop an approach for using Shadow CAT with a variable test-length termination rule and to evaluate the feasibility of the approach in operational CAT programs. This study has three research objectives:

1. Construct linear models for Shadow CAT using variable-length termination rules;
2. Evaluate the approach by assessing the extent to which all content constraints are met and equal measurement precision is achieved; and
3. Evaluate the performance of the approach in comparison with WDM and MPI content balancing methods.

Content Balancing Methods

Three content balancing methods were included in this study: WDM, MPI, and shadow test approach. The first two are heuristic methods and Shadow CAT is based on linear programming. The WDM and MPI methods are well known and widely used. A full list of heuristic methods was not included in this study because the focus of the study was to extend Shadow CAT to variable-length termination rules. A brief description of each method is given below.

WDM

The WDM (Stocking & Swanson, 1993) works by calculating the weighted sum of two components: the deviation from the content targets and the deviation from the item information target

for each available item. The item with the smallest sum is selected and administered to the examinee. To be more specific, the target is to minimize

$$\sum_{j=1}^J w_j d_{L_j} + \sum_{j=1}^J w_j d_{U_j} + w_\theta d_\theta, \quad (1)$$

subject to

$$\sum_{i=1}^N g_{ij} x_i + d_{L_j} - e_{L_j} = L_j, j = 1, \dots, J, \quad (2)$$

$$\sum_{i=1}^N g_{ij} x_i + d_{U_j} - e_{U_j} = U_j, j = 1, \dots, J, \quad (3)$$

$$\sum_{i=1}^N I_i(\theta) x_i + d_\theta - e_\theta = \infty, \quad (4)$$

$$d_{U_j}, d_{L_j}, e_{U_j}, e_{L_j} \geq 0, j = 1, \dots, J,$$

$$d_\theta, e_\theta \geq 0$$

$$x_i \in \{0, 1\}, i = 1, \dots, N,$$

where J denotes the number of constraints; N denotes the total number of items; w_j is the weight for constraint j ; L_j and U_j represent the lower and upper bounds of constraint j ; d_{L_j} and d_{U_j} denote the deficit from the lower and upper bounds of constraint j ; and e_{L_j} and e_{U_j} represent the excess from the lower and upper bounds of constraint j . Similarly, d_θ and e_θ denote the deficit and excess from the target test information. And g_{ij} equals to 1 if item i has property j and equals 0 if item i does not have property j . The decision variable x_i is a binary variable that either equals to 1 if item i is selected or equals to 0 if item i is not selected.

MPI Method

This method was proposed by Cheng and Chang (2009) with the goal of achieving fewer constraint violations and better exposure control than WDM. It works by calculating the priority index for each available item in the pool, and selecting the item with the largest index as the next item to administer. The priority index is defined as in Equation 5:

$$PI_i = I_i \prod_{j=1}^J (w_j f_j)^{c_{ij}}. \quad (5)$$

A constraint relevancy matrix $\mathbf{C}$, of dimension $N \times J$, is constructed, with $c_{ij} = 1$ if constraint j is relevant to item i and 0 otherwise. I_i denotes the Fisher information of item i at the current ability estimate and w_j is the weight assigned to each constraint j . The two-phase item selection framework (Cheng & Chang, 2009) focuses on satisfying the lower bound requirements in the first phase and meeting the upper bound requirements in the second phase. After x_j items have been selected for constraint j , the scaled “quota left” f_j is defined in Equation 6 for Phase 1 and defined in Equation 7 for Phase 2:

$$f_j = \frac{L_j - x_j}{L_j}, \quad (6)$$

$$f_j = \frac{U_j - x_j}{U_j}. \quad (7)$$

Shadow Test Approach to CAT

Most CAT approaches select items for administration directly from the item pool. In contrast, a typical Shadow CAT works in two steps: (a) from the item pool, assembles a complete form called shadow test, which satisfies all the statistical and nonstatistical constraints, includes all previously administered items, and maximizes the test information given the current ability estimate and (b) selects the optimal item from the free items, that is, items not selected and administered to the students, in the shadow test. As each step is optimal, the final test is optimal. Shadow CAT will be successful in avoiding constraint violations, which cannot be guaranteed when applying heuristic methods. The process of assembling the shadow is handled by ATA and it uses mathematical programming techniques. A mixed-integer programming (MIP) solver is needed to find the optimal solutions for the mathematical model used in the ATA. There are commercial solvers available, such as Xpress, Gurobi, OPL-CPLEX 6.3, and LINGO 12.0 (LINDO), and freely available solvers, such as lp_solve version 5.5 (Diao & van der Linden, 2011; Konis, 2009). An example of a basic ATA model can be found in van der Linden and Diao (2011).

Termination Rules

Two variable-length termination rules were used in this study, namely, the *SE* termination rule and the *PSER* termination rule. *SPRT* and *GLR* termination rules are for computerized classification tests and were not included in this study. Under the usual regularity conditions, the inverse of Fisher information is asymptotically equal to the variance of the ability parameter. So asymptotically, *SE* and *MFI* termination rules should be equivalent based on their definitions. However, with a limited number of items, the inverse of Fisher information is only an approximate estimate of the variance of the ability parameter. Therefore, in adaptive testing with a limited number of items, it is difficult, if not impossible, to set the thresholds for *SE* and *MFI* termination for direct comparisons between them. As a result, this study only included the *SE* and *PSER* termination rules. The *SE* and *PSER* rules are briefly described below. Detailed descriptions of each rule can be found in Weiss and Kingsbury (1984) and Choi et al. (2011).

Standard Error Termination Rule

A CAT will end when the *SE* of an examinee's ability estimate, $\hat{\theta}$, falls below a prespecified value upon updating $\hat{\theta}$. In this study, the probability of a correct response to item i is modeled by the three-parameter logistic model (3PLM),

$$P_i(\theta) = c_i + (1 - c_i) \frac{\exp[a_i(\theta - b_i)]}{1 + \exp[a_i(\theta - b_i)]}, \quad (8)$$

where a_i , b_i , and c_i are the discrimination, difficulty, and pseudo-guessing parameters, respectively. The Fisher information value for item i is computed as follows:

$$I_i(\theta) = a_i^2 \frac{(1 - P_i)}{P_i} \left[\frac{P_i - c_i}{1 - c_i} \right]^2. \quad (9)$$

And the Fisher information value for the k administered items is as follows:

$$I(\theta) = \sum_{i=1}^k I_i(\theta). \quad (10)$$

The formula for computing SE of $\hat{\theta}_n$ after n items under asymptotic assumptions is shown below:

$$SE(\hat{\theta}_n) \approx \frac{1}{\sqrt{I(\hat{\theta}_n)}}. \quad (11)$$

According to the fixed SE termination rule, if the calculated SE value is smaller than the prespecified SE threshold, then the test is terminated. In addition to Fisher information, observed information is another choice for the SE computation in real testing.

PSER Termination Rule

The PSER rule examines whether the SE would be reduced significantly by administering additional items. The PSER index is defined as

$$PSER = SE(\hat{\theta}_n) - SE_{\text{pred}}(\hat{\theta}_n), \quad (12)$$

and

$$SE_{\text{pred}}(\hat{\theta}_n) = SE(\hat{\theta}_{n+1}^{(1)}) \cdot P(u_{n+1} = 1 | \hat{\theta}_n) + SE(\hat{\theta}_{n+1}^{(0)}) \cdot P(u_{n+1} = 0 | \hat{\theta}_n),$$

where $P(u_{n+1} = 1 | \hat{\theta}_n)$ and $P(u_{n+1} = 0 | \hat{\theta}_n)$ are the probabilities of correct and incorrect responses to the next item conditioning on the current estimate $\hat{\theta}_n$, $\hat{\theta}_{n+1}^{(1)}$ and $\hat{\theta}_{n+1}^{(0)}$ are the corresponding updated estimate of θ , and $SE(\hat{\theta}_{n+1}^{(1)})$ and $SE(\hat{\theta}_{n+1}^{(0)})$ are the corresponding SE based on the updated estimate of θ . Conceptually, the PSER termination rule combines the SE and MFI termination rules. In addition to the SE threshold, the PSER termination rule defines two additional thresholds, referred to as the Hypo parameter (H^-) and the Hyper parameter (H^+). The method works by comparing four conditions. If the SE of the current ability estimate, $SE(\hat{\theta}_n)$, is greater than the prespecified SE value, the authors checked whether the PSER index is greater than the prespecified H^- . If PSER is greater than H^- , it means that there are some items in the pool which can reduce the SE of ability estimate by more than H^- , so then the test continues; otherwise, the test will be terminated. On the contrary, if $SE(\hat{\theta}_n)$ is less than the prespecified SE value, then check whether the PSER index is greater than the prespecified H^+ . If PSER is greater than H^+ , it means that there are more items in the pool that can further reduce the SE of the ability estimate significantly by more than H^+ , and so then the test continues; otherwise, the test will be terminated. This approach balances measurement precision and item utilization.

Shadow Test Approach for Constructing Variable-Length Tests

For practical reasons, most CAT applications set lower and upper boundaries on the test length in addition to the termination rule to avoid unexpectedly short/long tests. When constructing

Shadow CAT with variable-length termination rules, the lower boundary is set to ensure that all minimum content constraints could be satisfied within the first stage. This allows the test to freely terminate in the second stage.

There are two stages in constructing a Shadow CAT with variable-length termination rules. In the first stage, shadow tests are constructed using the minimum test length as a fixed test-length constraint. That is, each shadow test in the first stage has the same test length and the test length is the minimum test length required for the CAT. The authors use n to represent number of items administered to the examinee and $\hat{\theta}_n$ to represent the current ability estimate. The basic steps of the first stage are as follows:

1. Assemble a shadow test based on the current ability estimate $\hat{\theta}_n$, which includes all n items that have been administered, satisfies all constraints, and maximizes test information based on $\hat{\theta}_n$.
2. Select the item with the maximum Fisher's information value from the free items in the shadow test, and administer the item to the examinee as the $(n + 1)$ th item.
3. After the examinee gives a response, update the ability estimate to $\hat{\theta}_{n+1}$.
4. Repeat Steps 1 to 3 until the number of items administered reaches the minimum test length.

After the test reaches the minimum test-length requirement, a new test construction method is used for the second stage of the test. Namely, the maximum test length is used as the fixed test-length constraint and it is check whether the test should be stopped according to the termination rule. If the test reaches the maximum test length without reaching the termination rule criteria, the test will be terminated. Detailed steps are given below for both the *SE* termination rule and PSER termination rules in the second stage (i.e., when n is larger than the minimum test length and smaller than the maximum test length).

In conjunction with the *SE* termination rule, the second stage of the Shadow CAT approach works by calculating the $SE(\hat{\theta}_n)$ and comparing it with the prespecified *SE* criterion.

1. If the $SE(\hat{\theta}_n)$ is greater than the criterion, the test continues by repeating the Steps 1 to 3 in the first stage and constructing the next shadow test based on $\hat{\theta}_n$.
2. If the $SE(\hat{\theta}_n)$ is smaller than the criterion, the test terminates.

When using Shadow CAT together with the PSER termination rule, the following steps are taken.

1. Assemble a shadow based on $\hat{\theta}_n$, and select the item with the maximum Fisher Information value from the free items in this shadow.
2. Calculate the $SE(\hat{\theta}_n)$ and compute the PSER based on the selected $(n + 1)$ th item.
3. Compare the $SE(\hat{\theta}_n)$ and the PSER value with the following prespecified criteria:
 - i. If the $SE(\hat{\theta}_n)$ is less than the prespecified *SE* value and
 - a. If $PSER \geq H^+$, the test continues; administer the selected $(n + 1)$ th item to the examinee.
 - b. Otherwise, the test is terminated.
 - ii. If the $SE(\hat{\theta}_n)$ is greater than the prespecified value and
 - a. If $PSER \geq H^-$, the test continues; administer the selected $(n + 1)$ th item to the examinee.
 - b. Otherwise, CAT is terminated.

Table 1. Descriptive Statistics of the Item Pool.

Item statistics	<i>a</i>	<i>b</i>	<i>c</i>
<i>M</i>	1.617	0.000	0.189
<i>SD</i>	0.769	1.309	0.064
Minimum	0.150	-3.791	0.037
Maximum	3.084	4.928	0.431

A primary difference between the *SE* and PSER termination rules is that the *SE* termination rule checks the *SE* of the ability estimate first in determining whether to continue the test; if the current *SE* is not below the specified threshold, the shadow test is constructed for selecting the next item to administer. In contrast, the PSER termination rule constructs the shadow test first to calculate the PSER of the next item to administer. If the test continues, the next item is administered without constructing another shadow test.

Simulation

Item Pool

The item pool consisted of 165 items from a large-scale formative assessment program calibrated under the 3PLM based on responses from more than 5,000 students. Descriptive statistics of the item response theory (IRT) parameter estimates are given in Table 1. The constraints imposed by the operational program are summarized in Table 2. In total, there were 27 content-based constraints (54 when considering lower/upper bounds) and corresponding weights for the heuristic approaches (WDM and MPI). All constraints in Table 2 required the number of items from each content category to be between the corresponding lower and upper bounds.

Simulation Setup

The simulation study was split into the following two cases. The first case examined the degree to which each method met the test constraints. The second case obtained a conditional sample, which was used to evaluate the measurement precision of each method. In Case 1, a sample of 1,000 simulated examinees was drawn from a standard normal distribution. In Case 2, tests were replicated 500 times at each of $-3.0(0.5)3.0$ points. The minimum and maximum test lengths were 24 and 36, respectively. α -stratification (Chang & Ying, 1999) was used as the exposure control method. The item pool was divided into three strata according to the *a*-parameter values. Each stratum included one third of the items in the pool. The first eight items of the CAT could only be selected from the stratum with the lowest *a* parameters; the next eight items could only be selected from the stratum with moderate *a* parameters. All remaining items in the CAT could be selected from the entire pool (i.e., from all three strata).

The adaptive tests used expected a posteriori (EAP) as the interim ability estimates and maximum likelihood estimates as the final ability estimates. The prior distribution was taken to be the standard normal.

In total, six scenarios were simulated with combinations of two termination rules, that is, *SE* and PSER, and three content balancing methods, that is, WDM, MPI, and Shadow CAT. Each combination was simulated with both Case 1 and Case 2 samples. For termination rules, the prespecified *SE* criterion was set equal to 0.2. For the PSER method, the prespecified *SE* criterion was set equal to 0.2, H^+ equal to 0.03, and H^- equal to 0.01. The choice of the threshold

Table 2. Constraints and Associated Weights of the Simulated CAT.

ID	Lower bound	Upper bound	Weight
C01	24	36	20
C02	16	36	20
C03	0	10	20
C04	13	22	20
C05	2	5	20
C06	3	6	20
C07	6	6	20
C08	8	16	18
C09	3	8	18
C10	3	8	18
C11	4	8	18
C12	3	6	18
C13	0	0	18
C14	0	0	23
C15	0	0	23
C16	0	0	23
C17	0	0	23
C18	0	0	23
C19	0	5	20
C20	0	10	20
C21	0	0	23
C22	0	0	23
C23	0	5	20
C24	0	2	20
C25	0	2	20
C26	0	10	20
C27	0	10	20

Note. CAT = computerized adaptive testing.

values was somewhat arbitrary, albeit typical in the literature (Choi et al., 2011). In general, the cutoff values should be selected on a case-by-case basis. When a researcher is dealing with a new bank, it is recommended that a series of simulation studies be conducted with different cutoff values to decide the optimal ones that serve the best to the new bank and the test.

Evaluation Criteria

Constraint violation was assessed by calculating the total number of times the constraints were met out of 1,000 simulees, and the average constraint value over 1,000 simulees. Measurement precision was assessed by calculating the overall bias, root mean square error (RMSE), mean standard error (*SE*) of the final estimates, and the standard deviation (*SD*) of the final *SE*. Average test length was also calculated.

Results

Case 1 Results

Case 1 examined constraint satisfaction. Tables 3 and 4 report the measures of constraint satisfaction for the content balancing methods for each termination rule. Across the six conditions, Shadow CAT met the test constraints for all 1,000 examinees. The MPI method also met the

Table 3. Number of Times Constraints Met and Average Values of the Constraints With SE Termination Rule.

ID	Shadow		WDM		MPI	
	Number meet	Average value	Number meet	Average value	Number meet	Average value
C1	1,000	27.4	1,000	34.4	1,000	32.9
C2	1,000	20.9	1,000	30.2	1,000	29.1
C3	1,000	6.5	1,000	4.2	1,000	3.8
C4	1,000	14.8	1,000	20.6	1,000	18.4
C5	1,000	2.7	1,000	3.7	1,000	3.5
C6	1,000	3.9	1,000	4.2	1,000	5.1
C7	1,000	6.0	1,000	6.0	1,000	6.0
C8	1,000	10.1	1,000	12.9	1,000	13.1
C9	1,000	4.9	1,000	6.4	1,000	6.1
C10	1,000	4.2	1,000	4.3	1,000	4.7
C11	1,000	4.8	1,000	6.8	1,000	5.3
C12	1,000	3.4	1,000	4.0	1,000	3.9
C13	1,000	0.0	1,000	0.0	1,000	0.0
C14	1,000	0.0	1,000	0.0	1,000	0.0
C15	1,000	0.0	1,000	0.0	1,000	0.0
C16	1,000	0.0	1,000	0.0	1,000	0.0
C17	1,000	0.0	1,000	0.0	1,000	0.0
C18	1,000	0.0	1,000	0.0	1,000	0.0
C19	1,000	1.3	1,000	1.7	1,000	2.5
C20	1,000	0.5	1,000	2.2	1,000	2.5
C21	1,000	0.0	1,000	0.0	1,000	0.0
C22	1,000	0.0	1,000	0.0	1,000	0.0
C23	1,000	0.0	1,000	0.0	1,000	0.0
C24	1,000	0.0	1,000	1.0	1,000	0.6
C25	1,000	1.1	1,000	1.8	1,000	1.8
C26	1,000	8.9	991	9.9	1,000	9.9
C27	1,000	9.7	992	9.8	1,000	9.9

Note. WDM = weighted deviation method; MPI = maximum priority index.

constraint requirements for both *SE* and PSER termination rules. The WDM method generally met the constraints when using the *SE* termination rule, albeit with two exceptions. Out of 1,000 examinees, nine of the tests did not meet constraint C26, and eight of the tests did not meet constraint C27. The WDM method generally met the constraints when using the PSER termination rule, albeit with one exception. Out of 1,000 examinees, 20 of the tests did not meet constraint C11. Because only a small number of examinees using WDM method did not meet the constraints, the average constraint values of those constraints were similar to the ones from Shadow CAT and MPI methods.

The measures of precision for the Case 1 results are summarized in Table 5. The Shadow CAT had the smallest bias, RMSE, final *SE*, and *SD* of the final *SE* across both termination rules. Between the two heuristic methods, MPI outperformed WDM for most of the measures.

Case 2 Results

Case 2 used conditional samples to evaluate the measurement precision of each method. Figure 1 shows bias, RMSE, *SE* of the final estimates, and *SD* of *SE* of the final estimates for all the methods. The first row of the graph shows that Shadow CAT had the smallest bias of all the

Table 4. Number of Times Constraints Met and Average Values of the Constraints With PSER Termination Rule.

ID	Shadow		WMD		MPI	
	Number meet	Average value	Number meet	Average value	Number meet	Average value
C1	1,000	24.2	1,000	24.2	1,000	24.0
C2	1,000	17.9	1,000	22.0	1,000	21.5
C3	1,000	6.3	1,000	2.2	1,000	2.5
C4	1,000	13.1	1,000	13.2	1,000	13.0
C5	1,000	2.0	1,000	2.0	1,000	2.0
C6	1,000	3.1	1,000	3.0	1,000	3.0
C7	1,000	6.0	1,000	6.0	1,000	6.0
C8	1,000	8.7	1,000	9.0	1,000	8.7
C9	1,000	3.9	1,000	4.3	1,000	4.1
C10	1,000	3.8	1,000	3.3	1,000	3.8
C11	1,000	4.7	980	4.4	1,000	4.3
C12	1,000	3.2	1,000	3.2	1,000	3.1
C13	1,000	0.0	1,000	0.0	1,000	0.0
C14	1,000	0.0	1,000	0.0	1,000	0.0
C15	1,000	0.0	1,000	0.0	1,000	0.0
C16	1,000	0.0	1,000	0.0	1,000	0.0
C17	1,000	0.0	1,000	0.0	1,000	0.0
C18	1,000	0.0	1,000	0.0	1,000	0.0
C19	1,000	1.3	1,000	0.0	1,000	1.5
C20	1,000	0.4	1,000	0.0	1,000	0.5
C21	1,000	0.0	1,000	0.0	1,000	0.0
C22	1,000	0.0	1,000	0.0	1,000	0.0
C23	1,000	0.0	1,000	0.0	1,000	0.0
C24	1,000	0.0	1,000	0.0	1,000	0.0
C25	1,000	1.0	1,000	0.0	1,000	0.7
C26	1,000	8.7	1,000	4.2	1,000	8.6
C27	1,000	9.6	1,000	5.4	1,000	2.0

Note. PSER = predictive standard error reduction; WDM = weighted deviation method; MPI = maximum priority index.

Table 5. Bias, RMSE, Final SE, and SD of Final SE for All Conditions Based on Case I Results.

Termination rule	Content balancing method	Bias	RMSE	Final SE	SD of final SE
SE	Shadow	−0.007	0.311	0.255	0.421
	MPI	−0.017	0.393	0.309	0.623
	WDM	−0.028	0.419	0.316	0.581
PSER	Shadow	−0.007	0.302	0.256	0.356
	MPI	−0.017	0.415	0.335	0.655
	WDM	−0.019	0.492	0.408	0.619

Note. RMSE = root mean square error; WDM = weighted deviation method; MPI = maximum priority index; PSER = predictive standard error reduction.

content balancing methods, regardless of the termination rule; MPI had smaller bias than WDM for both termination rules. The second row shows that Shadow CAT had the smallest mean square errors (MSEs), especially at the extreme true ability points. The pattern was consistent across both termination rules. The third and fourth rows of the graph show the *SE* of the final

Figure 1. Bias, RMSE, final SE of estimates, and SD of final SE of estimates for all three content balancing methods under different stopping rules.
Note. RMSE = root mean square error; PSER = predictive standard error reduction; CAT = computerized adaptive testing; WDM = weighted deviation method; MPI = maximum priority index.

Figure 2. Average test length for all three content balancing methods under different stopping rules. Note. PSER = predictive standard error reduction; CAT = computerized adaptive testing; WDM = weighted deviation method; MPI = maximum priority index.

estimates and *SD* of the *SE* of the final estimates. Shadow CAT outperformed WDM and MPI in these measures as well, and the pattern was consistent across both termination rules. Between WDM and MPI, MPI had better measurement precision.

Figure 2 shows the average test length for the three content balancing methods. The test length indicated the efficiency of the method. Namely, the method with the shortest average test length was the most efficient method because the method was able to meet the same test requirements as the other methods, albeit with fewer items.

The first panel shows that Shadow CAT had the shortest average test length when the *SE* termination rule was used, whereas WDM had the longest average test length. The PSER termination rule is similar to *SE*, but can override the *SE* criterion in favor of improving measurement precision through the administration of additional items. When PSER termination rule was used, the three content balancing methods had similar average test lengths except for the extreme ability values, where MPI had shorter test lengths. The average test lengths of MPI were the same as the minimum test length set by the simulation.

The total number of items used in the simulation for each method as an index of item utilization was also counted. For PSER termination, the numbers are 135, 97, and 75 for the Shadow CAT, MPI, and WDM methods, respectively. For *SE* termination, the numbers are 136, 105, and 103 for the Shadow CAT, MPI, and WDM methods, respectively.

In terms of constraint satisfaction, the three methods performed similarly to Case 1. Out of 6,500 examinees, both Shadow CAT and MPI had no violations. WDM had violations to two constraints when combined with the *SE* termination rule and had violations to one constraint when combined with the PSER termination rule.

Conclusion and Discussion

This study had three goals. The first and most important one was to construct Shadow CAT with a variable-length termination rule. Variable-length termination is commonly used in CAT

and can achieve better measurement efficiency than fixed-length termination. Yet no previous research attempted to combine Shadow CAT with variable-length termination rules. This study demonstrated that variable-length Shadow CAT can be constructed and implemented in operational programs.

The second goal of the study was to examine constraint satisfaction and measurement precision under the variable-length approach. In the simulation study, the Shadow CAT method was shown to meet all test constraints, which is noteworthy considering the complexity of the constraints. The measurement precision of the examinee scores was similar, although the scores at the extreme ends of the scale were measured with less precision. The reduced measurement precision was a byproduct of item pool quality. Namely, there were not enough items at the extreme ends of the scale to support sound measurement of the corresponding ability scores. Thus, to better utilize adaptive testing, that is, to truly tailor the test to each individual student, a sound item pool design is needed (He & Dia0, 2014).

The third goal of the study was to compare the performance of Shadow CAT with WDM and MPI. The results of the simulation study showed that both Shadow CAT and MPI satisfied all constraint requirements, whereas the WDM method had some violations. In terms of measurement precision, Shadow CAT consistently outperformed WDM and MPI across all criteria, including bias, MSE, *SE* of final estimates, *SD* of *SE* of final estimates. In most cases, Shadow CAT also had better item utilization. For the extreme cases, MPI had better item utilization when PSER termination rule was applied. The optimality of the Shadow CAT comes from the linear programming method it uses. The method uses a mathematical model to achieve optimal measurement precision while satisfying all test constraints. Namely, the linear programming technique selects the most informative test at a particular ability level, from all possible test forms that satisfy the constraints. In contrast, heuristic methods such as WDM and MPI endorse sequential item selection and attempt to balance the trade-off between measurement precision and constraint satisfaction by means of weights. For example, the observed failure of the WDM to meet all test constraints may have been caused by low weights for some of the constraints. However, if the constraint weights had been increased, the measurement precision could be compromised. Because extending Shadow CAT to variable test length is the focus of the study, only two content balancing methods were used for comparison. In the future, more content balancing approaches can be included for comparison.

The success of any CAT depends largely on the quality of the item pool. If the quality of the item pool is poor, Shadow CAT may run into infeasibility. And heuristic methods may have even more severe constraint violations. In the current study, the measurement precision of extreme abilities was poor across all methods, which was largely caused by the quality of the item pool. That is, there were not enough quality items in the pool to support the measurement of examinees with extreme abilities. In future studies, various item pools should be used as an additional condition to better evaluate the impact of the item pool.

In the simulation study, the minimum test length was derived to accommodate the lower bounds of various content constraints. However, from a measurement precision perspective, a smaller minimum test length should be set to better utilize the measurement efficiency provided by the variable-length termination rule. For example, if a smaller minimum test length had been set for the PSER termination rule, the authors may have observed more tests terminating before reaching 24 items. In future studies, it would be interesting to examine the relationship between the minimum test length specified and the actual test length observed. Maximum likelihood estimates were used as the final ability estimates, and it was noticed that in some cases (less than 1% of the sample), the final abilities did not converge for students with extreme true abilities. The authors would recommend adding using EAP as both the interim and final ability estimation method as an additional condition, especially when the item pool does not support

sufficient measurement for students with extreme abilities. In addition, it would be interesting to conduct future studies on comparing variable-length Shadow CAT with the fixed-length Shadow CAT. The studies need to take into consideration how to choose the test length for the fixed-length CAT so a fair comparison between variable-length CAT and fixed-length CAT can be conducted.

Content balancing is key to establishing validity, especially in educational measurement. Of the various approaches to content balancing currently available, Shadow CAT provides a flexible framework for adaptive testing solutions that require complex sets of constraints. This study showed how to construct Shadow CAT with two variable-length termination rules, *SE* and *PSER*, and demonstrated the superiority of the approach in comparison with two other popular heuristic methods, *WDM* and *MPI*. The new method extends the utility of the Shadow Test approach by allowing for the construction of variable-length tests. Namely, the new method allows for optimal test solutions to be delivered efficiently and/or with high measurement precision, depending on the needs of the testing program.

Acknowledgments

The authors thank the editor, Hua-Hua Chang, the associate editor, Chih-Hung Chang, and three anonymous reviewers for their helpful comments. The authors also thank Wim van der Linden, Seung Choi, and David King for their inputs.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

- Chang, H. H., & Ying, Z. (1999). a-Stratified multistage computerized adaptive testing. *Applied Psychological Measurement*, 23, 211-222.
- Cheng, Y., & Chang, H. H. (2009). The maximum priority index method for severely constrained item selection in computerized adaptive testing. *The British Journal of Mathematical and Statistical Psychology*, 62, 369-383.
- Choi, S. W., Grady, M. W., & Dodd, B. G. (2011). A new stopping rule for computerized adaptive testing. *Educational and Psychological Measurement*, 70, 37-53.
- Diao, Q., & van der Linden, W. J. (2011). Automated test assembly using lp_solve version 5.5 in R. *Applied Psychological Measurement*, 35, 398-409.
- Gialluca, K. A., & Weiss, D. J. (1979). *Efficiency of an adaptive inter-subset branching strategy in the measurement of classroom achievement* (Research Report 79-6). Minneapolis: Department of Psychology, Psychometric Methods Program, University of Minnesota.
- He, W., & Diao, Q. (2014, April). *Item pool design for CAT-review, demonstration, and future prospects*. Paper presented at the National Council on Measurement in Education, Philadelphia, PA.
- He, W., Diao, Q., & Hauser, C. (2014). A comparison of four item-selection methods for severely constrained CATs. *Educational and Psychological Measurement*, 74, 677-696.
- Konis, K. (2009). lpSolveAPI (Version 5.5.0.15) [Computer software]. Retrieved from <http://CRAN.R-project.org/package=lpSolveAPI>
- Luecht, R. M. (1998). Computer-assisted test assembly using optimization heuristics. *Applied Psychological Measurement*, 22, 224-236.

- Maurelli, V., & Weiss, D. J. (1981). *Factors influencing the psychometric characteristics of an adaptive testing strategy for test batteries* (Research Report 81-4). Minneapolis: Department of Psychology, Psychometric Methods Program, University of Minnesota.
- Patton, J. M., Diao, Q., & Boughton, K. (2013, April). *From paper-and-pencil to CAT: An application of mixed-integer programming*. Paper presented at the National Council on Measurement in Education, San Francisco, CA.
- Reckase, M. D. (1983). A procedure for decision making using tailored testing. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait theory and computerized adaptive testing* (pp. 237-254). New York, NY: Academic Press.
- Shin, C., Chien, Y., Way, W. D., & Swanson, L. (2009). *Weighted penalty model for content balancing in CATs*. Pearson. Retrieved from https://images.pearsonassessments.com/images/tmrs/tmrs_rg/WeightedPenaltyModel.pdf?WT.mc_id=TMRS_Weighted_Penalty_Model_for_Content_Balancing
- Stocking, M. L., & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement*, 17, 277-292.
- Thompson, N. A. (2009, June). *Utilizing the generalized likelihood ratio as a termination criterion*. Paper presented at the GMAC Invitational Conference on Computerized Adaptive Testing, Minneapolis, MN.
- van der Linden, W. J. (2005). *Linear models for optimal test design*. New York, NY: Springer.
- van der Linden, W. J., & Diao, Q. (2011). Automated test-form generation. *Journal of Educational Measurement*, 48, 206-222.
- van der Linden, W. J., & Glas, C. A. W. (Eds.). (2010). *Elements of adaptive testing*. New York, NY: Springer.
- van der Linden, W. J., & Reese, L. M. (1998). A model for optimal constrained adaptive testing. *Applied Psychological Measurement*, 22, 259-270.
- Wald, A. (1947). *Sequential analysis*. New York, NY: John Wiley.
- Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, 21, 361-375.