Weighted Penalty Model for Content Balancing in CATS

Chingwei David Shin
Yuehmei Chien
Walter Denny Way
Len Swanson

April 2009

PEARSON

*Using assessment
and research to
promote learning*

Abstract

This research report proposes a new model called the Weighted Penalty Model (WPM) for content balancing in computer adaptive testing. The WPM approach attempts to balance content properties across all content categories as well as other non-statistical constraints, while simultaneously considering item information at each item-selection level and the scarcity of items relative to some constraints. This is accomplished by assigning a penalty value to each eligible item in the item pool. An item will be deemed as more desirable for selection if 1) its penalty value is small and 2) it will not make any constraint violation when administered. The purpose of this study is to present the WPM approach and demonstrate its performance using simulation with real item pool data.

Introduction

Content balancing has been one of the biggest concerns in the implementation of computerized adaptive testing (CAT). Ideally, the selection of next items for a CAT should be based on the item information given the current estimate of proficiency, subject to the sometimes competing requirement of balancing the test content specifications. To date, some content balancing methods provide a way not only to balance content categories but also to balance other constraints, such as an overlap constraint, item set constraint, key distribution constraint, and others. The constraints mentioned above are sometimes called non-statistical constraints. In contrast with non-statistical constraints, the item information, item difficulty, and item discrimination are statistical constraints.

Several methods of content balancing have been developed and studied, such as the Constrained CAT (CCAT; Kingsbury & Zara, 1989), the Weighted Deviations Model (WDM; Stocking & Swanson, 1993) and the shadow test approach (STA; van der Linden, 2005). The CCAT, a straightforward method, looks for the content category for which the cumulative percentages of administered items currently is farthest below its target percentage. The WDM method, which is much more complicated, balances the constraints by weighting each constraint and computing the deviations from the desired test properties using binary programming. (Note that constraints and properties are used interchangeably in this study.) The STA selects items from a shadow test that is a linear test assembled prior to the selection of each item. The WDM and STA methods are similar in that both are based on projections of the future consequences of selecting an item. However, they differ in that the WDM calculates a projection of a weighted sum of the properties of the eventual test and the STA calculates a projection (shadow test) of a realization of the full test.

In this study, we propose a content balancing method called the Weighted Penalty Model (WPM). Based on an approach originally proposed by Segall and Davey (1995), this method attempts to balance content properties across all content categories as well as other non-statistical constraints. At the same time, it considers item information at each item-selection level and the scarcity of items relative to some constraints (that is, the degree to which items with properties associated with particular constraints are sufficiently represented in the pool). This is accomplished by assigning a penalty value to each eligible item in the item pool at each item-selection level. Items with smaller penalty values are deemed more desirable for selection. The penalty function used by WPM is an adjusted version of the original penalty function proposed by Segall and Davey (1995) and is referred to as the adjusted penalty function.

The purpose of this study is to introduce the WPM approach and demonstrate its performance using simulation with real item pool data and using empirical data from large-scale placement CAT.

Weighted Penalty Model

The WPM is implemented by forming a list of items to be candidates for the next item administered. The WPM involves three stages: 1) calculating the weighted penalty value for each eligible item in the pool; 2) assigning each eligible item into different groups (we refer to these as “color groups”); and 3) forming a list of candidate items. If an item exposure control method is used, one of the candidate items from the list is selected based on the specific item exposure control method. Otherwise, the first item in the list is selected to be administered.

Calculating the Weighted Penalty Value

The definitions and formulas for calculating the weighted penalty value are as follows:

Definitions: For each constraint *j*, define

$Upper_j$ as the upper bound of the proportion of items in the test that should have the property associated with constraint *j*;

$Lower_j$ as the lower bound of the proportion of items in the test that should have the property associated with constraint *j*;

Mid_j as the midpoint between $Upper_j$ and $Lower_j$; and

$Prevalence_j$ as the proportion of the items in the pool having the property associated with constraint *j*.

For example, if the test length is 20 items, the upper bound is 4 items, the lower bound is 0 items, the pool is 100 items, and 30 items in the pool have the property associated with constraint *j*, then $Upper_j = 0.2$ (4/20), $Lower_j = 0$ (0/20), $Mid_j = 0.1$, and $Prevalence_j = 0.3$ (30/100).

At any point in the test, to obtain the weighted penalty value for each item, the following steps are taken:

1. Compute *Prop*$_j$, which is the expected proportion of items with constraint *j* that will have been administered if all remaining items in the test are selected in proportion to their prevalence. That is,

$$Prop_j = (nadm_j + Prevalence_j \times nremaining) / testlength, \qquad (1)$$

   where *nadm*$_j$ is the number of items administered so far having this property, *nremaining* is the number of items remaining to be administered in the test (including this one), and *testlength* is the length of the test.

2. Compute *X*$_j$, which is the expected difference between *Prop*$_j$ and the constraint target, Mid_j, across the full length of the test. Thus,

$$X_j = (Prop_j - Mid_j). \qquad (2)$$

3. For each eligible item *i*, compute the penalty value for each constraint *j* using one of Equations (3) to (5) below:

$$P_{ij} = \left(\frac{1}{kD_j} X_j^2 + \frac{D_j}{k}\right) \times Z_{ij}, \text{ if } Prop_j < Lower_j, \qquad (3)$$

   where D_j is $Lower_j - Mid_j$, *k* is arbitrary but has been chosen to be 2, and Z_{ij} is 1 if item *i* has property *j*, otherwise Z_{ij} is 0.

$$P_{ij} = \left(\frac{1}{kA_j} X_j^2 + \frac{A_j}{k}\right) \times Z_{ij}, \text{ if } Prop_j \geq Upper_j, \qquad (4)$$

   where A_j is $Upper_j - Mid_j$ and, again, *k* is arbitrarily chosen to be 2.

$$P_{ij} = X_j \times Z_{ij}, \text{ if } Upper_j > Prop_j \geq Lower_j. \qquad (5)$$

4. For each item *i*, compute the total content penalty value that takes into account all the content constraints of item *i*:

$$F_i''' = \sum_{j=1}^{J} P_{ij} \times w_j, \qquad (6)$$

where w_j is the weight for constraint *j*.

5. Standardize the total content constraint penalty value:

$$F_i' = \frac{F_i''' - \min(F_i''')}{\max(F_i''') - \min(F_i''')}, \tag{7}$$

where $\min(F_i''')$ and $\max(F_i''')$ are the minimum and maximum F_i''' over all eligible items, respectively.

6. Given $\hat{\theta}$ as the current estimate of the ability for each item *i*, compute the standardized item information value:

$$SI_i(\hat{\theta}) = \frac{I_i(\hat{\theta})}{I_{max}(\hat{\theta})}, \tag{8}$$

where $I_i(\hat{\theta})$ is the information value of item *i* given $\hat{\theta}$ and $I_{max}(\hat{\theta})$ is the maximum information value across all eligible items given $\hat{\theta}$.

7. Compute the information penalty value that takes into account the information:

$$F_i'' = -SI_i(\hat{\theta})^2. \tag{9}$$

8. Finally, compute the weighted penalty value:

$$F_i = w' \times F_i' + w'' F_i'', \tag{10}$$

where w' and w'' are the weights for F' and F'', respectively.

The weights, w' and w'', referred to as the content constraint weight and the item information weight, respectively, can differ across the sequence of items selected. For each item selected, these weights act as “control parameters” (van der Linden, 2005), which control the trade-off between the content constraint and the item information. We have found it useful to set w'' using functions of the item sequence number in the test. The use of different functions results in various patterns for the relative weights of set w'' versus w'.

Figure 1 illustrates four different information weight patterns. Two of the patterns are based on a logistic function (*logistic* and *logistic+2*). In this pattern the information weight increases slowly in the beginning and at the end but rapidly in the middle. An additional two patterns are based on a quadratic function (*quadratic* and *quadratic+2*); in this pattern, the information weight increases slowly in the beginning and middle but rapidly at the end. The content constraint weight is set to be a constant value of 5. When the value of the content constraint weight is larger than the value of information weight, the CAT algorithm will tend to select the items that better fit the content constraints. When the value of the content constraint weight is smaller than the value of the information weight, the CAT algorithm will tend to select items that maximize information.

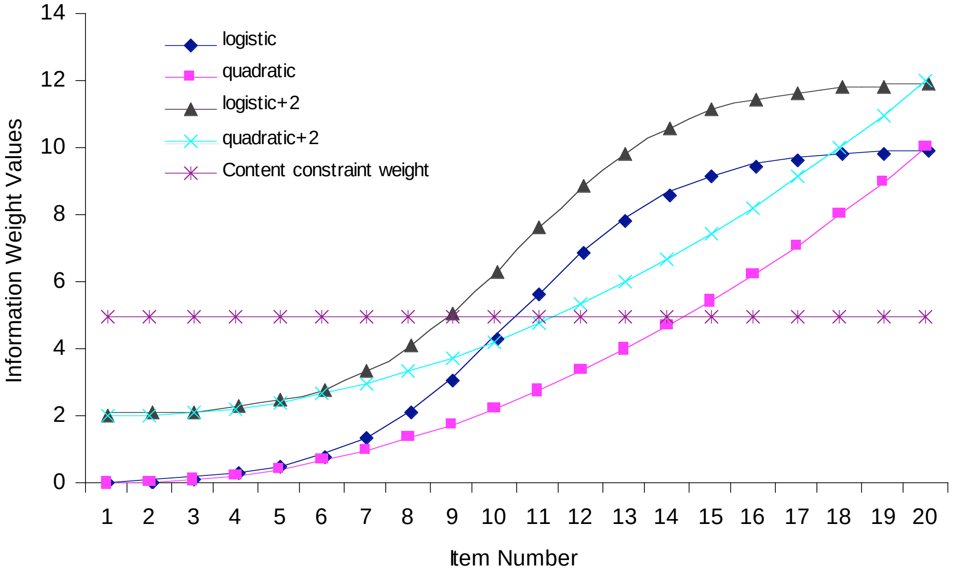

Figure 1. Alternate Information Weights as a Function of Item Sequence Number

These weights temper the sometimes undesirable characteristic of maximum information–based item-selection—that is, the tendency to choose only those items with the most desirable statistical properties. Employing a function that emphasizes content over information at the beginning of the adaptive test is an idea that is different in practice but similar in concept to the a-stratified item-selection method (Chang, Qian, & Ying, 2001). (In the a-stratified item-selection method, the items are stratified into a number of levels based on the *a*-parameters. The early stages of a test uses items with lower *a*'s and later stages use items with higher *a*'s.) The flexibility and ease with which w'' can be varied to change the relative emphasis of content and information is a distinct advantage of the WPM.

Assigning Items into Different Color Groups

As an item is associated with more than one constraint, some of the constraints associated with a specific item often are under their corresponding lower bounds (i.e., more desirable to be selected) while the other constraints associated with the same item are at or beyond their corresponding upper bounds (i.e., less desirable to be selected). For such items with both kinds of constraints (more desirable and less desirable), the algorithm likely would select one of them and cause the upper boundary violation. A grouping method was developed to avoid selecting an item that would cause any content violation while there are still other items that would not cause the same issue if selected.

In the grouping method, first, a flag is assigned to each of the constraints based on the number of items that have been administered so far and the constraints' upper and lower boundaries. To assign a flag to each of the constraints at each item-selection level, the following rules are used:

1. If the lower bound of the constraint has not been reached, "A" is assigned to this specific constraint;
2. If the lower bound of the constraint has been reached but not the upper bound, "B" is assigned to this specific constraint; and
3. If the upper bound has been either reached or exceeded, "C" is assigned to this specific constraint.

After all of the constraints have been assigned flags, each eligible item in the pool will be place in a color group based on the flags of its associated constraints. The rules are:

1. If the flags of the associated constraints for an item are all "A" or the combination of "A" and "B," this item is assigned to the "green" group;
2. If the flags of the associated constraints for an item are either the combination of "A," "B," and "C" or the combination of "A" and "C," this item is assigned to the "orange" group;

3. If the flags of the associated constraints for an item are all “B,” this item is assigned to the “yellow” group; and
4. If the flags of the associated constraints for an item are either the combination of “C” and “B” or all “C,” this item will be assigned to the “red” group.

Forming a List

After all the eligible items in the pool have been assigned to color groups, the list will be formed according to the following rules:

1. Between color groups, the order will be green, orange, yellow, and red; and
2. Within each color group, the items are ordered by the weighted penalty values from smallest to the largest.

Item-selection Procedure

After forming a list of items using WPM based on the item exposure control method used, one of the items from the list is selected to be administered. In this study, two item exposure control methods are adopted: the *Conditional Randomesque* (CR) method and the *Stocking and Lewis Conditional Multinomial* (SLCM; Stocking and Lewis, 2000) method.

The randomesque strategy (Kingsbury and Zara, 1989) randomly selects the next item to be administered from the group of the most informative items, given the current estimated theta, where the group size is predetermined (e.g. 2, 3, 4,…10). The CR method in this study is the variation of the regular randomesque strategy. Given the current estimated theta, the CR method selects the next item from a group of items, where group size is predetermined for that ability range. (For example, 3, 4, 4, 5, 4, 3 are the group sizes for the 6 theta ranges if the whole theta scale is divided into 6 ranges.) The rest of items in that group that are not selected will be blocked from further item selection. In doing so, the CR strategy allows a preset maximum exposure rate for each ability range to be stipulated and provides a reasonable assurance that the maximum exposure rate will be constrained to that level.

The SLCM strategy directly controls the item exposure rate conditional on estimated theta by deriving an exposure-control parameter for each item at each ability level. The exposure-control parameter is valued from 0 to 1. Once the list of items has been formed using the WPM method, the first k items from the list are selected for further consideration.

To randomly select one item from the k items, first, a cumulative multinomial distribution is formed based on the exposure-control parameters of those k items. Then, a random number is generated. The corresponding item in the cumulative multinomial distribution based on the random number is selected to be administered. All items preceding the one administered will be blocked from item selection for the rest of the test. (See Stocking and Lewis, 2000, for more details.) The value of k is predetermined based on the pool size and the test length.

In addition to the two item exposure control methods, another item-selection method adopted in this study directly selects the first item from the list without applying any item exposure control method, which is referred to as the none_IEC method in this study.

Simulation Study

In this section, WPM is demonstrated through simulation using a real item pool from a large-scale placement CAT. The study factors, data, simulation design, and evaluation criteria are described in the following sections.

Study Factors

The factors investigated in this study include three item-selection methods as mentioned previously and two item information weight patterns, which results in a total of six study conditions.

For the item information weight—w'' in Equation (10)—at each item-selection level, two different information weight patterns, *Logistic+2* and *Quadratic+2,* are studied. Figure 2 shows the two patterns for a 12-item test. The content constraint weight is set to be a constant value of 5. When the value of the content constraint weight is larger than the value of information weight, the CAT algorithm might tend to select the items that better fit the content constraint. When the value of content constraint weight is smaller than the value of information weight, the CAT algorithm might tend to select items that provide more information.

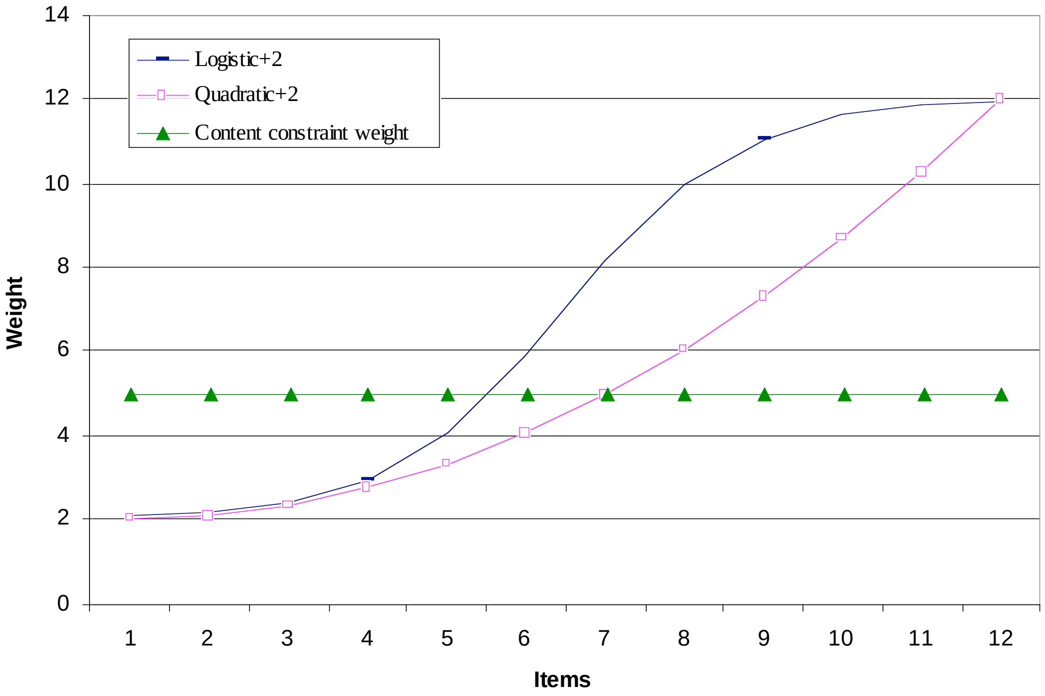

Figure 2. Information weight curves

Data

The item pool contained 565 items. The mean of the discrimination parameters (a) was 1.163, with SD = 0.486. The mean of the difficulty parameter (b) was 0.044, with SD=1.100. The mean of the guessing parameter (c) was 0.198, with SD=0.092. Note that items could be indicated as overlapping with other items; once one item within an overlap set was administered, the other items in the set were blocked from further consideration. For this item pool, one item might be either in multiple overlap groups or in none at all.

The length of the CAT is fixed at 12 items. There are 21 constraints as shown in Table 1 for the 12-item CAT. The first 17 constraints are content constraints and the last four are key distributions constraints.

Table 1. Constraints.

Constraint	Weight	Lower	Upper
C1	5	0	1
C2	15	1	1
C3	10	0	1
C4	10	0	1
C5	20	1	1
C6	15	1	1
C7	15	1	1
C8	5	0	1
C9	15	1	1
C10	15	1	1
C11	15	1	2
C12	5	0	1
C13	5	0	1
C14	10	1	1
C15	5	0	1
C16	5	0	2
C17	5	0	1
C18	0.5	2	5
C19	0.5	2	5
C20	0.5	2	5
C21	0.5	2	5

Simulation Design

In this study, the ability values are estimated using maximum likelihood estimation. The maximum and minimum theta points were 5.0 and -5.0, respectively. For both the SLCM and CR methods, ability scale was divided into 10 theta ranges by 9 cut points, which were -1.483, -0.865, -0.479, -0.165, 0.12, 0.395, 0.679, 1.003, and 1.449. Two thousand simulees were generated in this study from a normal distribution with mean -0.59 and SD 1.37, calculated from the empirical sample distribution of the large-scale placement CAT. The predefined group sizes for the CR method are all 4 for the 10 theta ranges in order to control the maximum exposure rate around 0.25. The SLCM exposure-control parameters are generated through simulation with *k*=15 and desired maximum exposure rate 0.25.

Evaluation Criteria

Five criteria were used to assess the performance of the WPM approach:

(1) Overall bias for theta estimation,

$$\text{Bias} = \sum_{j=1}^{N} (\hat{\theta}_j - \theta_j) \Big/ N \,, \qquad (11)$$

where N is the number of simulees, θ_j and $\hat{\theta}_j$ are the true and estimated theta for simulee j, respectively;

(2) The mean square error (MSE) for theta estimation,

$$\text{MSE} = \sum_{j=1}^{N} (\hat{\theta}_j - \theta_j)^2 \Big/ N \,; \qquad (12)$$

(3) The correlation between true theta and estimated theta;

(4) The average conditional standard error of measurement (CSEM); and

(5) The percentage of tests that matches the target property.

Results

Table 2 lists the results of the simulation study. It shows that the WPM method worked very well on content balancing for each study condition with the on-target rate nearly or equal to 100% This indicates that the WPM method handled content balancing very well for the two information weight patterns—Logistic+2 and Quadratic+2—with or without item exposure control methods used. Within each item exposure control method, the results regarding measurement precision obtained from using Logistic+2 are comparable to those obtained from using Quadratic+2. Therefore, the two different information weight patterns yielded similar results for this studied CAT design with the real item pool.

Under each study condition, the bias value is very small. However, a certain amount of precision loss is expected with any of the item exposure control methods. As expected, the none_IEC method had better measurement precision in terms of smaller MSE, higher correlation between true and estimated thetas, and smaller CSEM as compared to the other two IEC methods. For either of the information weight patterns, the CR method had slightly higher correlation values and slightly smaller MSE values than the SLCM method. On the contrary, the SLCM method had slightly smaller CSEM values.

Table 2. Results of Simulation

ITEM EXPOSURE CONTROL METHOD	None_IEC		CR		SLCM	
INFO WEIGHT VECTOR	Logistic+2	Quadratic+2	Logistic+2	Quadratic+2	Logistic+2	Quadratic+2
Bias	0.0096	0.0163	0.0139	0.0010	0.0000	0.0093
MSE	0.3218	0.3320	0.4446	0.4413	0.5148	0.4809
Correlation	0.9242	0.9208	0.9046	0.9059	0.8889	0.8965
CSEM	0.3291	0.3306	0.3647	0.3663	0.3626	0.3629
ON-TARGET RATE	99.70%	100%	99.90%	99.95%	99.85%	99.95%

Table 3 lists the results with respect to item exposure control. In terms of evaluating item exposure control results, maximum item exposure rates and the pool usage were calculated for the six study conditions. For the none_IEC method, the maximum item exposure rate was about .5 for both of the information weight patterns, which means for one out of every two students would see this specific item. The maximum item exposure rate for the CR method was 0.2575 for both of the information weight patterns, which was close to the preset level—0.25. For the SLCM method, the maximum item exposure rate was about 0.21 for both of the information weight patterns, which is 4% below the preset level. For this study, the SLCM method had best fit to the preset maximum exposure rate when the same level was set for both the CR and the SLCM methods.

The pool usage was expressed through the distribution of the item usage rate. Note that the average item usage rate is 2.12%, based on a 12-item test with 565 items in the pool (12 divided by 565 is about 0.212). When none of the item exposure control methods was applied, nearly 70% of the items in the pool were not used and about 0.7% of the items had the item usage rate beyond 30%. Within either of the IEC methods, the two information weight patterns had similar results for the distribution of the item usage rates. The CR method significantly reduced the zero item usage rate from 70% to 18% and increased the number of items into the two item usage categories—0%~2% and 2% to 10%. The SLCM method reduced the zero item usage rate from 70% to 33%, and also increased the number of items into the two categories—0%~2% and 2% to 10%. The CR method had better performance regarding the zero item usage rate in this study.

Table 3. Results of Item Exposure Control

ITEM EXPLOSURE CONTROL METHOD	**None_IEC**		**CR**		**SLCM**	
INFO WEIGHT VECTOR	**Logistic+2**	**Quadratic+2**	**Logistic+2**	**Quadratic+2**	**Logistic+2**	**Quadratic+2**
Max_IE	0.4860	0.5100	0.2575	0.2575	0.2120	0.2150
ITEM USAGE 0%	69.56%	69.20%	18.23%	17.52%	32.74%	33.45%
ITEM USAGE 0 ~ 2%	13.63%	13.81%	48.50%	48.14%	32.39%	30.80%
ITEM USAGE 2%~ 10%	7.79%	8.67%	30.97%	31.68%	32.04%	32.92%
ITEM USAGE 10% ~ 20%	6.55%	5.84%	1.42%	1.77%	2.65%	2.65%
ITEM USAGE 20% ~ 30%	1.77%	1.77%	0.88%	0.88%	0.18%	0.18%
ITEM USAGE 30% and UP	0.70%	0.71%	0	0	0	0

Empirical Data Analyses

The WPM has been adopted by a large-scale CAT program. This large-scale CAT program has the same CAT design as the simulation study in this pape. The initial empirical data of 1,066 examinees were available from this large-scale CAT program. The empirical CAT used *Quadratic+2* as the information weight vector and CR method for item exposure control. The values for the information weight vector are: 2.06, 2.083, 2.311, 2.744, 3.322, 4.066, 4.975, 6.05, 7.289, 8.694, 10.264, and 12.

The empirical data analysis results are presented in Table 4. In Table 4, C1 to C21 were 21 constraints, and lower and upper presents the lower and upper limits of the number of items in a CAT associated with that constraint in the first column. The on-target rate shows the percentage of examinees whose tests meet that specific constraint. The six columns on the right side of Table 4 present the percentage of examinees who have the specific number of items (0 to 5) associated with the constraint in the first column. For example, there should be at least 1 item and at most 2 items in a CAT that have the constraint C11. For those 87 examinees, 5.75% of examinees had 1 item associated with C11 and 94.25% of examinees had 2 items associated with C11; therefore, all examinees (100%) had reached the targeted number of items (either 1 or 2) associated with C11.

The results show the on-target rates were 100% for all constraints, which indicated the WPM method worked well for this specific CAT design.

Table 4. Empirical Data Analysis Results

				Cross tabulation for Percent of Examinees and Number of Items in the Constraints					
Constraint	Lower	Upper	On_Target_Rate	0	1	2	3	4	5
C1	0	1	100.00%	44.83%	55.17%	0.00%	0.00%	0.00%	0.00%
C2	1	1	100.00%	0.00%	100.00%	0.00%	0.00%	0.00%	0.00%
C3	0	1	100.00%	26.44%	73.56%	0.00%	0.00%	0.00%	0.00%
C4	0	1	100.00%	14.94%	85.06%	0.00%	0.00%	0.00%	0.00%
C5	1	1	100.00%	0.00%	100.00%	0.00%	0.00%	0.00%	0.00%
C6	1	1	100.00%	0.00%	100.00%	0.00%	0.00%	0.00%	0.00%
C7	1	1	100.00%	0.00%	100.00%	0.00%	0.00%	0.00%	0.00%
C8	0	1	100.00%	68.97%	31.03%	0.00%	0.00%	0.00%	0.00%
C9	1	1	100.00%	0.00%	100.00%	0.00%	0.00%	0.00%	0.00%
C10	1	1	100.00%	0.00%	100.00%	0.00%	0.00%	0.00%	0.00%
C11	1	2	100.00%	0.00%	5.75%	94.25%	0.00%	0.00%	0.00%
C12	0	1	100.00%	78.16%	21.84%	0.00%	0.00%	0.00%	0.00%
C13	0	1	100.00%	86.21%	13.79%	0.00%	0.00%	0.00%	0.00%
C14	1	1	100.00%	0.00%	100.00%	0.00%	0.00%	0.00%	0.00%
C15	0	1	100.00%	74.71%	25.29%	0.00%	0.00%	0.00%	0.00%
C16	0	2	100.00%	26.44%	73.56%	0.00%	0.00%	0.00%	0.00%
C17	0	1	100.00%	11.49%	88.51%	0.00%	0.00%	0.00%	0.00%
C18	2	5	100.00%	0.00%	0.00%	24.14%	45.98%	24.14%	5.75%
C19	2	5	100.00%	0.00%	0.00%	19.54%	57.47%	19.54%	3.45%
C20	2	5	100.00%	0.00%	0.00%	25.29%	55.17%	16.09%	3.45%
C21	2	5	100.00%	0.00%	0.00%	28.74%	59.77%	10.34%	1.15%

Conclusion

In summary, the WPM method handled the content balancing very well in this study, with or without applying the item exposure control methods for both the empirical and simulated data. The IEC results indicated that although the SLCM method provided lower maximum item exposure rates, the CR method had great utility for increasing pool utilization in this study. As expected, both of the IEC methods had a certain amount of precision loss compared with the none_IEC method.

References

Chang, H., Qian, J., & Ying, Z. (2001). a-Stratified multistage Computerized Adaptive Testing with *b* blocking. *Applied Psychological Measurement, 25*(4), 333-341.

Kingsbury, G., Zara, A. (1989). Procedures for Selecting Items for Computerized Adaptive Tests. *Applied Measurement in Education, 2(4),* 359-75.

Segall, D. O. & Davey, T. C. (1995, June). *Some New Methods for Content Balancing Adaptive Tests.* Paper presented at the annual meeting of the Psychometric Society, Minneapolis MN.

Stocking, M.L. & Lewis, C. (2000). *Methods of Controlling the Exposure of Items in CAT* in W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized Adaptive Testing: Theory and Practice,* (pp. 163–182). Norwell MA: Kluwer.

Stocking, M. L., & Swanson, L. (1993). A Method for Severely Constrained Item Selection in Adaptive Testing. *Applied Psychological Measurement, 17,* 277-292.

Van der Linden, W. J. (2005). A Comparison of Item-Selection Methods for Adaptive Tests with Content Constraints. *Journal of Educational Measurement. 42(3),* 283-302.