A Comparison of Four Item-Selection Methods for Severely Constrained CATs

Wei He

Northwest Evaluation Association (NWEA)

Qi Diao

CTB/McGraw-Hill

Carl Hauser

Northwest Evaluation Association (NWEA)

Paper presented at the 2013 NCME annual meeting. Please address all questions to Wei He wei.he@nwea.org.

A COMPARISON OF FOUR ITEM-SELECTION METHODS FOR SEVERELY CONSTRAINED CATS

ABSTRACT

This study compares the four existing procedures handling the item selection in severely constrained computerized adaptive tests (CAT). These procedures include weighted deviation model (WDM), weighted penalty model (WPM), maximum priority index (MPI), and shadow test approach (STA). Severely constrained CAT refer to those adaptive tests seeking to meet a complex set of content constraints simultaneously, acknowledging that an item usually carries multiple attributes that are inclusive to each other. In addition, two modified versions of the MPI procedure are introduced to deal with the situation in which the priority indices for all eligible items are zero. Given the item pool characteristic and the adaptive model within which this study is conducted, the results indicate that the shadow test approach, among all candidate methods, works the best in terms of measurement accuracy and constraint management, except that it makes the poorest use of items. All heuristic approaches do not differ significantly from each other in terms of measurement accuracy and constraint management at the lower bound level. However, the WPM method appears to perform considerably better in overall constraint management than both WDM and MPI methods. Regarding the two modified MPI procedures, the M2_MPI (i.e., the one assuming "move at its own pace" for each constraint) appears to perform better than the M1_MPI (i.e., the one assuming "move at the same pace" for all constraints) in overall constraint management. Regarding the three variations of WPM procedure, the WPM_fixed (1), i.e., the one adopting different weights to calculate content and item information penalty values, works better than other two variations. Limitations and further research directions are also discussed.

A COMPARISON OF FOUR ITEM-SELECTION METHODS FOR SEVERELY CONSTRAINED CATS

INTRODUCTION

Test specifications lay out rules for including items in a test (Swanson & Stocking, 1993). These rules typically consist of a series of constraints—both statistical and non-statistical—on item properties/attributes. Examples of statistical (i.e., psychometric) constraints might include target item or test information function while examples of non-statistical (i.e., non-psychometric) constraints include content specifications, item format, depth of knowledge, or key location (see van der Linden & Boekkooi-Timminga (1989) and Swanson & Stocking (1993) for an extensive description of possible constraints). In a standardized testing program, it is imperative that test forms meet the same non-statistical specifications across individual examinees and provide reliable ability estimates. For computerized adaptive tests (CATs), this requirement can only be met by forcing the item selection algorithm to combine the objective of maximizing information with a strategy that can impose the same set of non-statistical specifications on the items selected for administration (van der Linden, 2005).

Several approaches, known as content balancing in the CAT literature, have been proposed to manage the non-statistical constraints while at the same time assembling a CAT that can measure efficiently and accurately. These approaches include Kingsbury and Zara's (1989) constrained CAT (CCAT) method, the weighted deviations model (WDM; Stocking & Swanson, 1993), the shadow-test approach (STA; van der Linden & Reese, 1998), the modified multinomial model (MMM; Chen & Ankenmann, 2004), the modified CCAT (MCCAT; Leung, Chang, & Hau, 2003), the two-phase item selection procedure for flexible content balancing method (Cheng, Chang, & Yi, 2007), the weighted penalty model (WPM; Shin, Chien, Way & Swanson, 2009), and the maximum priority index (MPI) method (Cheng & Chang, 2009). Of all the above named methods, the CCAT, the MCCAT, and the MMM can be viewed as stemming from a common methodological approach in that an item pool is partitioned into several sub-pools by key item attributes and items are spirally selected from across sub-pools to meet pre-determined content specifications. In general, these methods are more appropriate for use in CATs in which only a single item attribute is considered in item selection, which is usually the one used to partition the item pool) or attributes are considered to be mutually exclusive to one another. In contrast, the STA, the WDM, the WPM, and the MPI stem from different methodological approach that seeks to meet CAT specifications calling for items to be selected that meet a complex set of constraints simultaneously, acknowledging that an item usually carries multiple attributes that are inclusive to each other. This study is focused on these four methods of handling a complex set of constraints.

These four methods have some similarities and differences. For example, the STA is a mathematical programming method whereas the other three are heuristic methods. The STA, employs a constrained sequential optimization approach that treats test specifications as constraints that must be imposed on the item selection. As such, the STA can guarantee perfect adherence to test specifications as long as the solutions are feasible. The other three methods, however, treat test specifications as objectives, in which specifications are formulated as constraints and treated as goal values. The tests are assembled to be optimal with respect to these goal values. As long as the tests are optimal, it is not a problem that the some bounds are violated. This explains why, for the WDM, the WPM, and the MPI, feasible solutions can be ensured but with less certainty that test specifications can be met (Cheng & Chang, 2009; Robin et al., 2005; van der Linden, 2005). The STA selects an item for administration by solving a sequence of simultaneous optimization problems, thus being computationally intensive and potentially requiring a trade-off between the search speed and optimality of its solution or even infeasible solutions. With the increasing computer power and availability of extremely powerful commercial solvers, such as CPLEX 12.1 (IBM ILOG), however, this may not be a problem. In comparison, the three heuristic methods use a sequential heuristic search method by selecting the next item that can minimize the weighted sum of deviations from the target value (i.e., the WDM), minimize the weighted values (i.e., the WPM), or maximize the priority index (i.e., the MPI), thus always ensuring a feasible solution. Unlike the STA, the applications of MPI, the WPM and the WDM require input (from content experts) about weights and potentially a considerable amount of time to adjust the heuristic. For example, it may be time consuming to find the best weights, in order for the tests to achieve the best trade-off among constraints as well as the best trade-off between measurement quality and constraint management.

Several studies (Cheng & Chang, 2009; Moyer, Galindo, & Dodd, 2012; Robin et al., 2005; Shin, Chien, & Way, 2012; van der Linden, 2005) have been conducted to compare the performance of some of the methods discussed above. However, these studies are not conclusive with respect to which procedure performs better. Moreover, none of the studies have focused on these four methods at the same time. As such, this study is undertaken to compare the effectiveness of these four approaches in handling item selection in a severely constrained CAT, i.e., a CAT with a complex set of constraints.

BRIEF INTRODUCTIONS TO FOUR DIFFERENT ITEM SELECTION METHODS

Below are brief descriptions of how each method works. For detailed information, the reader is referred to the original papers.

Maximum Priority Index (MPI)

The MPI requires that a priority index, PI, be calculated for each eligible item *j* in the item pool at each item selection step. Items with larger priority index values are deemed more desirable for administration. A two-phase item selection framework (Cheng et al., 2007) is proposed to implement the MPI method in the presence of flexible content balancing, i.e., constraints are specified in the form of both lower and upper bounds.

$$PI_j = I_j \prod_{k=1}^{K} (w_k f_k)^{c_{j_k}}$$

where I_j represents item Fisher information of item *j* at the provisional ability estimate, and w_k represents the weight assigned to constraint *k*. c_{jk} represents the constraint relevancy matrix with 1 indicating an item has constraint *k* and is zero otherwise. The relevancy index matrix is usually identified before hand by content experts. f_k represents the scaled “quota left” (Cheng & Chang, 2009) of constraint *k*. In the first phase which is focused on meeting the lower bound requirements, f_k is calculated by $f_k = \frac{(l_k - x_k)}{l_k}$. Once all lower bounds, l_k, are met, then the test moves to the second phase with focus on meeting the upper bound, u_k, requirements and f_k is calculated by replacing l_k with u_k in the second phase. x_k indicates the number of items carrying constraint *k* that have been administered. In the first phase of item selection, when constraint *k* reaches its lower bound, f_k is set to zero which results in a priority index value of zero. This treatment serves the purpose of setting those "fulfilled" constraints to be "dormant thus waiting for other content areas to catch up.

However, since an item's priority index value is a product of item information and the weighted "quota left" for each constraint evaluated against whether an item carries a specified property, an undesirable side effect can result, i.e., the MPI values for all eligible items being zero. When this situation occurs, item selection can't proceed as the MPI method intends, or is reduced to random item selection in some sense. To handle this situation, two modified approaches are proposed in this study. In the first approach (denoted as *M1_MPI*), if the situation occurs in which the PI value for all eligible items is zero in the first phase, then for each individual constraint *k* in item *j*, f_k is assigned a value that is much lower than the minimum non-zero value calculated for all constraints. When the lower bounds are reached at the end of the first phase, then the test moves to the second phase.

In the second approach (denoted as *M2_MPI*), whenever a constraint *k* reaches its lower bound, the test moves to its upper bound. This approach allows each constraint to "move at its own pace." When the upper bound for a constraint is satisfied, its PI value becomes zero meaning no more items can be selected out of this content category. If the situation occurs in which the PI values for all eligible items are zero, then for each individual constraint *k* in item *j*, f_k is assigned a value that is much lower than the minimum non-zero value calculated for all constraints.

Weighted Deviation Model (WDM)

The weighted deviation model (WDM) was originally developed by Stocking and Swanson (1993) out of the concern about possible poor-quality item pools in large-scale test assembly. In CAT, the WDM works through three major steps: 1) for every item not already in the test, the deviations from the content targets and from the target item information value are calculated respectively as if the items were added to the test; 2) for each item, calculating its weighted sum across all constraints; and 3) items with the smallest sum are selected for administration.

The WDM is formulated as the follows:

Minimize

$$\sum_{k=1}^{K} W_k d_{L_k} + \sum_{k=1}^{K} W_k d_{U_k} + W_{\theta} d_{\theta}$$

Subject to

$$\sum_{i=1}^{N} g_{ik} x_i + d_{L_k} - e_{L_k} = L_k \quad k = 1, \ldots, K$$

for the lower bound. And

$$\sum_{i=1}^{N} g_{ik} x_i + d_{U_k} - e_{U_k} = U_k \quad k = 1, \ldots, K$$

for the upper bound

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

$$d_{U_k}, d_{L_k}, e_{L_k}, e_{U_k} \geq 0 \quad k = 1, \ldots, K$$

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

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

where N denotes the number of items in the item pool, *k* denotes the number of constraints, W_k denotes the weight assigned to each constraint, L_k and U_k denote the lower and upper bound for each *k* constraint respectively, d_{L_k} and d_{U_k} denote the deficit from the lower bound and surplus from upper bound respectively, e_{L_k} and e_{U_k} denote excess from lower bound and deficit from upper bound respectively, and d_{θ} denote the “deviations” from target item information. W_{θ} needs to be defined. g_{ik} is 1 if item *i* has property *k* and 0 otherwise. x_i is a binary decision variable: it equals 1 if i^{th} item is included in the test and 0 otherwise.

Weighted Penalty Model (WPM)

The original WPM method was proposed by Segall and Davey (1995) and later was modified by Shin et al. (2009). This method operates through three major steps: 1) calculating a penalty value to each eligible item in the item pool at each item selection level, 2) assigning each eligible item into different groups (referred to as “color groups” in Shin et al., 2009) based on how well the specified constraints are represented by the items administrated so far relative to the properties of items eligible for selection, and 3) selecting items with smaller penalty value for administration according to certain priority criteria. The overall penalty value F_i for an individual item is a weighted sum of penalty values calculated for content and item information.

$$F_i = w'*F_i^{'} + w''*F_i^{''}$$

$F_i^{'}$ and $F_i^{''}$ are content and item information penalty values. The weights w' and w'', represent content constraint and item information, respectively, and can take any values in theory. They act as “control parameters”; the values they are assigned control the trade-off between the content constraint and item information. This is an advantage that WPM has over other methods.

Shadow Test Approach (STA)

The STA, originally proposed by van der Linden and Reese (1998), is based on the ideas developed for the application of linear programming to optimal test assembly. The key point that makes the STA different from other approaches is that items are not selected directly from the pool but from a shadow test, i.e., a full-size test assembled prior to the administration of each item in the adaptive test. Unlike the other heuristic approaches, the statistical information from the test items at the current ability estimate are viewed as the objective function to be optimized and all other specifications are treated as constraints within which which the optimization has to take place (van der Linden, 1998; van der Linden, Ariel, & Veldkamp, 2006; Veldkamp & van der Linden, 2000).

In CAT, the STA works by the following major procedures: 1) initialize the estimator of the ability parameter; 2) assemble the first shadow test that meets all constraints and optimizes the objective function; 3) administer an eligible item in the shadow test that can provide maximum information at the current ability estimate; 4) Update the parameters in the test assembly model; 5) assemble a new shadow test but fix items already administered; and 6) repeat Steps 2-5 until expected test length is reached.

METHODS

Item Pool

This simulation study is carried out with a retired item pool from a large-scale operational CAT program. This item pool contains 361 items calibrated according to the three-parameter item response theory logistic model (3PL IRT). Descriptive statistics of this item pool are described below in Table 1. Figure 1 presents an overlay of item information for all items in the pool. It can be seen that the test is expected to provide better measurement accuracy and precision within ability levels ranging between -2.5 and .5 than others.

[Insert Table 1 about here]

[Insert Figure 1 about here]

Table 2 lays out the constraints, their associated weights, and their lower and upper bounds. The item attributes fall into five major constraint categories and each category is further divided into 2-5 finer areas. There are 18 constraints in total, each associated with a weight. Lower and upper bounds specify the number of selected items having specific attributes. Besides, there is one additional constraint, i.e., conflicting items, that requires selecting a certain item excludes the selection of other item(s).

[Insert Table 2 about here]

CAT Specifications

The CAT algorithm mimics that of an operational large-scale CAT program. The IRT model is the 3PL IRT logistic model. To obtain the current ability estimate before both correct and incorrect responses are available, the *expected a posteriori* (EAP) method is used with a normal $N(0,1)$ prior. Once both correct and incorrect responses are obtained, the maximum likelihood estimation (MLE) method is used. The test length is set as 20. The initial ability is set as 0 for all individual simulees. The exposure control procedure (except in maximum item information and random item selection methods) is the 5-4-3-2-1 randomesque method (McBride & Martin, 1983). That is, the first item is selected out of a group of best five candidate items identified by the candidate item selection method, the second item is out of the a group of best four candidate items, so on and so forth.

The constrained item selection methods, i.e., STA, WDM, WPM, and MPI, are adopted to select the items that can meet the specifications described in Table 2. For the MPI method, the modified MPI methods (i.e., M1_MPI and M2_MPI) are implemented since the trial simulation using the original MPI method encountered the condition that priority indices for all eligible items are zero. For the WPM, as the selection of the weights for content penalty and item information penalty can affect item selection, three versions of the WPM methods are implemented, i.e., 1) WPM_fixed(1) in which w' and w'' are set unequally as 6 and 2 respectively; 2) WPM_fixed(2) in which w' and w'' are set equally as 5; and 3) WPM_flx in which w' is set as 6 but w'' is set as a logistic function of the item sequence number (see Figure 2) as suggested in Shin et al. (2009). The purpose of the third version is to focus selection on items that better fit the content constraints at the early stage of the test and on items that can provide higher item information at the later stage of test respectively. In addition, two other item selection methods are adopted. The first one is maximum item information (denoted as *MI*) in which at each item selection step, the item that can provide the maximum information is administered regardless of any constraints. The second item selection method is random item selection (denoted as *RAND*) in which an item is randomly selected for administration regardless of any constraints. The *MI* and *RAND* methods are included to serve as the baseline measures for measurement precision and exposure control. In total, nine item selection methods are considered in this study.

[Insert Figure 2 about here]

Simulation Design

A random sample of 10,000 simulees drawn from the standard normal distribution is administered the CATs described above. In addition, conditional samples are used. Specifically, 3,000 examinees with abilities conditioning at each ability point from -3 to 3 at an increment of .5 are considered, resulting in 39,000 (i.e., 3,000*13) examinees. Both overall and conditional samples are administered the adaptive tests described above.

Evaluation Criteria

Two sets of statistics are computed for both overall and conditional samples. For overall samples, these statistics included:

1) Measure accuracy and precision. Overall bias, absolute bias (Abs(bias)), mean square error (MSE), and fidelity coefficient $r_{\theta,\hat{\theta}}$ (i.e., the correlation between estimated and true thetas) are computed as follows:

$$Bias = \frac{1}{N}\sum_{i=1}^{N}(\hat{\theta}_i - \theta_i)$$

$$Abs(bias) = \frac{1}{N}\sum_{i=1}^{N}|(\hat{\theta}_i - \theta_i)|$$

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

$$\rho_{\theta,\hat{\theta}} = \frac{\sum_{i=1}^{N}(\theta_i - \bar{\theta})(\hat{\theta}_i - \bar{\hat{\theta}})}{s_\theta s_{\hat{\theta}}}$$

where N is the number of examinees, $\hat{\theta}_i$ and θ_i are the estimated and true ability of the i^{th} examinee, and $s_{\hat{\theta}}$ and s_θ are the standard deviations of the estimated and true abilities, respectively.

2) Constraint violation. Several perspectives can be used to capture and shed light on the performance of different item selection methods with regard to constraint management: 1) the percentage of tests that violate the overall content management [denoted by *v_overall(%)*] which includes any test with content attribute distributions that fall out of the intended lower and upper bounds; 2) the average number of violated constraints across all examinees; and 3) the percentage of tests that violate the intended lower bound [denoted by *v_lower(%)*]. Violations of the intended lower bound can be treated as a less restrictive criterion than violations of the upper bound. It can be argued that the lower bound usually represents the minimum threshold that a test is expected to meet. That is, a test can be considered as valid in contents as long as the minimum criteria are fulfilled.

3) Test security. The maximum item exposure rate, the number of overexposed items (i.e., items with exposure rate above .2), and test overlap rate, defined as the

percentage of items in common to the two test events of two randomly selected examinees, are reported for each of the four item selection methods.

4) Item pool utilization. Two indices are computed: 1) skewness of item exposure rate distribution (χ^2; Chang & Ying, 1999) and 2) the number of underexposed items (i.e., items that have never been administered).

$$\chi^2 = \sum_{i=1}^{n} \frac{(r_i - L/n)^2}{L/n}$$

where r_i is the observed exposure rate for the i^{th} item, L is the test length, and n is the total number of items in the pool. This χ^2 index can measure the departure of an item's actual exposure from uniform item exposure and thus quantify the efficiency of item pool usage.

For each conditional sample, conditional bias and conditional MSE are computed. In addition, constraint violation is reported including the proportion of test events that violate the lower bounds and the proportion of test events that violate the overall content management.

RESULTS

Overall Sample

Table 3 reports the summary statistics for measurement quality, exposure control, and item usage across all nine item selection methods. In comparison to the MI method, all candidate item selection methods in question produce comparable yet negligible magnitudes of biases—both average and absolute. WPM_fixed(2), STA, and M2_MPI produce errors (i.e., MSE) slightly closer to the MI than the remaining methods. In terms of item exposure control, WDM overexposes the largest number of items and is followed by STA and WPM_flex. The two variations of WPM methods, which calculate penalty value with constants for content and item information, give the smallest number of overexposed items. In general, all candidate item selection methods do not seem to yield significantly different number of overexposed items from the MI method. The reason can be attributed to the "5-4-3-2-1" exposure control method which basically selects the "best" item after the fourth item. Regarding item usage, WPM_fixed(1), WDM, and WPM_flex appear to make better use of items than the others, and STA makes the poorest use of items by using less than half of items and producing the highest chi-square value.

[Insert Table 3 about here]

Table 4 presents constraint violation information across all nine item selection methods from different perspectives. It is obvious that, among all candidate selection methods, the STA best controls test constraints with zero violations both at overall and lower bound levels. Among all heuristic methods (i.e., WDM, MPI, and WPM), the WPM methods (all three variations) outperform other two heuristic methods with respect to overall violation and average number of violations, and the WPM_fixed (1) manages the content constraints much better than other two variations of the WPM method by having only 1.8% percent of test events with constrait violations. In comparison with the WPM method, both WDM and MPI methods witness constraint violation at overall level for almost all test events. However, across the three heuristic-based methods, very few test events, if any, violate lower bounds; the WDM does not yield any tests violating lower bounds. Two variations of WPM methods (i.e., WPM_fixed (2) and WPM_flex) result in some test events with the number of violations up to 7. For WDM and M2_MPI, the largest number of violations is up to 3. Figure 3 reports the proportions of tests with different numbers of constraint violations across all item selection methods.

[Insert Table 4 about here]

[Insert Figure 3 about here]

Conditional Samples

Figures 4 and 5 portray the conditional bias and MSE across different ability levels. In general, all candidate item selection methods procedure produce comparable measurement accuracy and precision within the ability range between -2 and 1.5. This is consistent with the overlay of item information curves in Figure 1.

[Insert Figure 4 and Figure 5 about here]

With respect to constraint management, STA ranks the best without any violations both at lower bounds and overall levels. As Figure 6 indicates, like STA, WDM does not witness any tests violating lower bounds. Two variations of MPI methods and WPM_fixed (1) also do a good job in managing the lower bound, though with trivial violations. Both WPM_fixed (2) and WPM_flex give significant proportions of tests violating lower bounds in particular within the ability range between -3 and 0. When it comes to the overall constraint management, as Figure 7 suggests, WPM, in particular, WPM_fixed(1), outperforms two other heuristic methods, i.e., WDM and MPI.

[Insert Figure 6 and Figure 7about here]

SUMMARY AND DISCUSSION

This study compares the four existing procedures handling the item selection in severely constrained CATs. In addition, two modified versions are introduced based on the MPI procedure to deal with the situation in which the priority indices for all eligible items are zero. The shadow test approach, among all candidate methods, works the best in terms of measurement accuracy and constraint management, except that it makes the poorest use of items. All heuristic approaches do not differ significantly from each other in terms of measurement accuracy and constraint management at the lower bound level. However, the WPM method appears to perform considerably better in overall constraint management than both WDM and MPI methods given the item pool used in this study. Regarding the two modified MPI procedures, the M2_MPI (i.e., the one assuming “move at its own pace” for each constraint) appears to perform better than the M1_MPI (i.e., the one assuming “move at the same pace” for all constraints) in overall constraint management. Regarding the three variations of WPM procedure, the WPM_fixed (1), i.e., the one adopting different weights to calculate content and item information penalty values, works better than other two variations.

As STA is a mathematical method, it is not surprising that it performs better than heuristic methods, in particular, in terms of constraint management. Unlike the heuristic methods, which are very easy to implement, the STA needs a linear programming solver to be able to work. Among all three heuristic methods, the MPI is more “static” than others in that it does not require much fine-tuning while being implemented. Both WDM and WPM, in particular, WPM, requires certain level of fine-tuning to achieve the best trade-off between measurement quality and constraint management. The three variations of the WPM is an illustration of such a fine-tuning process and the results indicate how different the results can be given the choices of different weights. Unlike in Shin et al. (2009), the current study indicates that setting the weights in WPM as a constant throughout rather than a non-constant works better. This fine-tuning effort is called for whenever the underlying item pool changes or certain CAT component such as content exposure is changed. This should be something that testing programs need to be aware of before implementing these heuristic procedures.

Clearly, item pool characteristics play a role as to how these methods work. Though different heuristic methods perform differently in terms of constraint management given the current item pool, it is anticipated that they may work comparably given certain characteristics underlying the item pool. This is related to the area of optimal item pool design for CAT (see He, 2010; He & Reckase, 2011 for examples) and it can be an area that future studies can examine. At the same time, the exposure control procedure that the current study uses may be too loose, and future studies should consider adopting more stringent exposure controls. How these procedures work in the presence of testlets is also a topic that needs further exploration.

CATs are gaining wider use in the high-stakes educational achievement testing arena, in particular with the announcement from the Smarter Balanced Assessment Consortium that its tests will be administered in the form of CAT (“Smarter Balanced Assessment”, 2011). This initiative calls for more research in CAT topics specifically targeted for the educational testing arena. For real world high-stakes achievement tests, complex constraints are commonly imposed on the existing linear forms. It is expected that, as we transform to this new form of testing, these constraints, if not more, will be imposed on CATs. The results presented here not only provide better understanding on how each method works and how they perform relative to each other but also provide practical guidance into the implementation of different procedures.

REFERENCES

Chang, H. (2007). Book review: Linear models for optimal test design. *Psychometrika*, 72, 279-281.

Chen, S., & Ankenmann, R. D. (2004). Effects of practical constraints on item selection rules at the early stages of computerized adaptive testing. *Journal of Educational Measurement, 41*(2), 149-174.

Cheng, Y., & Chang, H. (2009). The maximum priority index method for severely constrained item selection in computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology*, 63, 369-383.

Cheng, Y., Chang, H., & Yi, Q. (2007). Two-phase item selection procedure for flexible content balancing in CAT. *Applied Psychological Measurement*, 31, 467-482.

He, W. (2010). *Optimal item pool design for a highly constrained computerized adaptive test*. Unpublished doctoral dissertaion. Michigan State University.

He. W., & Reckase, M. (2011). *Optimal item pool design for a highly constrained computerized adaptive test*. Paper presented at the National Council on Measurement in Education, Denver, CO.

Kingsbury, G. G., & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education, 2,* 359–375.

Leung, C. K., Chang, H., & Hau, K. (2003). Computerized adaptive testing: A comparison of three content balancing methods. *Journal of Technology, Learning, and Assessment*, 2(*5*).

McBride, J.R. & Martin, J.T. (1983). Reliability and validity of adaptive ability tests in a military setting. In D.J. Weiss (Ed.), *New horizons in testing* (pp. 223-226). New York: Academic Press.

Moyer, E., Galindo, J., & Dodd, B. (2012). Balancing flexible constraints and measurement precision in computerized adaptive testing. *Educational and Psychological measurement*. DOI: 10.1177/0013164411431838.

Robin, F., van der Linden, W. J., Eignor, D. R., Steffen, M., & Stocking, M. L. (2005). *A comparison of two procedures for constrained adaptive test construction* (ETS Research Rep No. RR-04-39). Princeton, NJ: Educational Testing Service.

Shin, C., Chien, Y., & Way, D. (2012). *A comparison of two content balancing methods for fixed and variable length computerized adaptive test*. Paper presented at the 2012 NCME annual conference, Vancouver, Canada.

Shin, C., Chien, Y., Way, W. D., & Swanson, L. (2009). *Weighted penalty model for content balancing in CATs*. Pearson. Retrieved from http://www.pearsonedmeasurement.com/downloads/research/Weighted%20Penalty%20Model.pdf.

Smarter Balanced Assessments. (2012). Retrieved July 26, 2012, from http://www.smarterbalanced.org/smarter-balanced-assessments/.

Stocking, M. L., & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement, 17,* 277–292.

Swanson, L., & Stocking, M. L. (1983). A model and heuristic for solving very large item selection problems. *Applied Psychological Measurement, 17*, 151-166.

van der Linden, W. J. (2005). A comparison of item-selection methods for adaptive tests with content constraints. *Journal of Educational Measurement, 42,* 283-302.

van der Linden, W. J., & Boekkooi-Timminga, E. (1989). A maximin model for test design with practical constraints. *Psychometrika, 54*, 237-248.

van der Linden, W. J., & Reese, L. (1998). A model for optimal constrained adaptive testing. *Applied Psychological Measurement, 22* (3), 259-270.

APPENDIX

Table 1. *Descriptive Statistics of the Item Pool*

Item Statistics	*a*	*b*	*c*
Mean	0.975	-0.729	0.149
SD	0.389	0.882	0.084
Max	2.125	3.617	0.500
Min	0.022	-2.730	0.007

Table 2. *Constraints and Associated Weights*

Constraint Category	Constraint Code	Weight	Lower Bound	Upper Bound
I	C1	10	10	10
	C2	10	10	10
II	C3	10	6	8
	C4	10	6	8
	C5	10	6	8
III	C6	5	2	5
	C7	5	5	8
	C8	5	2	5
	C9	5	2	5
	C10	5	2	5
IV	C11	11	0	1
	C12	11	0	1
	C13	10	0	1
	C14	10	0	1
V	C15	1	3	7
	C16	1	3	7
	C17	1	3	7
	C18	1	3	7

Table 3. *Summary statistics for measurement accuracy and precision, exposure control, and item usage for overall sample*

	WDM	M1_MPI	M2_MPI	WPM_fixed(1)	WPM_fixed(2)	WPM_flex	STA	MI	RAND
Bias	0.019	0.018	0.012	0.022	0.022	0.023	0.023	0.016	0.029
Abs(BIAS)	0.254	0.255	0.246	0.258	0.227	0.237	0.229	0.202	0.416
MSE	0.113	0.112	0.104	0.116	0.087	0.099	0.088	0.071	0.367
$r_{\theta,\hat{\theta}}$	0.948	0.948	0.951	0.947	0.959	0.954	0.959	0.967	0.862
% of overexposed items	10.80	9.14	9.14	8.59	8.86	9.70	9.97	11.36	0
Test overlap rate	32.31	35.50	34.77	35.68	36.13	32.00	37.02	36.02	5.56
% of never exposed items	37.95	37.40	43.21	14.40	47.92	31.02	54.02	64.54	0
Chi-square	96.65	108.16	105.55	108.82	110.45	95.41	111.95	110.07	0.11

Table 4. *Proportions of tests having different numbers of constraint violations, proportions of tests violating lower bounds only, proportions of tests having at least one violation, and average number of violations per test for overall sample*

No. of violations(V)	WDM	M1_MPI	M2_MPI	WPM_fixed(1)	WPM_fixed(2)	WPM_flex	STA	MI	RAND
0	0	0	0.22	98.2	75.09	66.32	100	0	0
1	44.96	18.16	76.89	1.78	16.71	20.38	0	0	0
2	54.75	59.07	22.44	0.02	3.06	8.21	0	0	0
3	0.29	20.51	0.45	0	2.99	2.09	0	0	0.02
4	0	2.14	0	0	1.4	1.79	0	0.01	0.40
5	0	0.12	0	0	0.57	0.87	0	0.03	1.73
6	0	0	0	0	0.16	0.28	0	0.46	5.90
7	0	0	0	0	0.02	0.05	0	1.89	13.61
8	0	0	0	0	0	0.01	0	5.89	21.77
9	0	0	0	0	0	0	0	19.77	23.53
10	0	0	0	0	0	0	0	15.84	17.41
11	0	0	0	0	0	0	0	18.8	10.08
12	0	0	0	0	0	0	0	16.7	3.99
13	0	0	0	0	0	0	0	16.49	1.20
14	0	0	0	0	0	0	0	4.11	0.30
15	0	0	0	0	0	0	0	0.01	0.06
$\overline{V}$	1.55	2.07	1.23	0.02	0.41	0.57		10.79	8.80
V_lower (%)	0	0.25	1.21	0.17	19.87	30.56	0	100	99.94
V_overall(%)	100	100	99.78	1.8	24.91	33.68	0	100	100

Figure 1. An overlay of item information

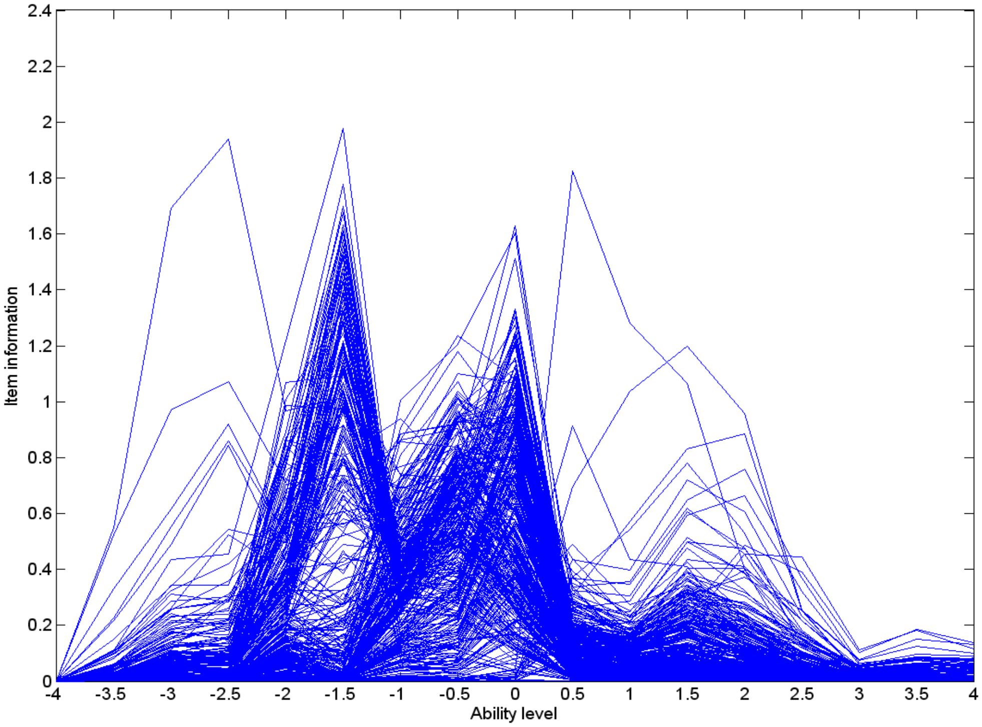


Figure 2. Weights used to compute the item information penalty value in WPM_flex

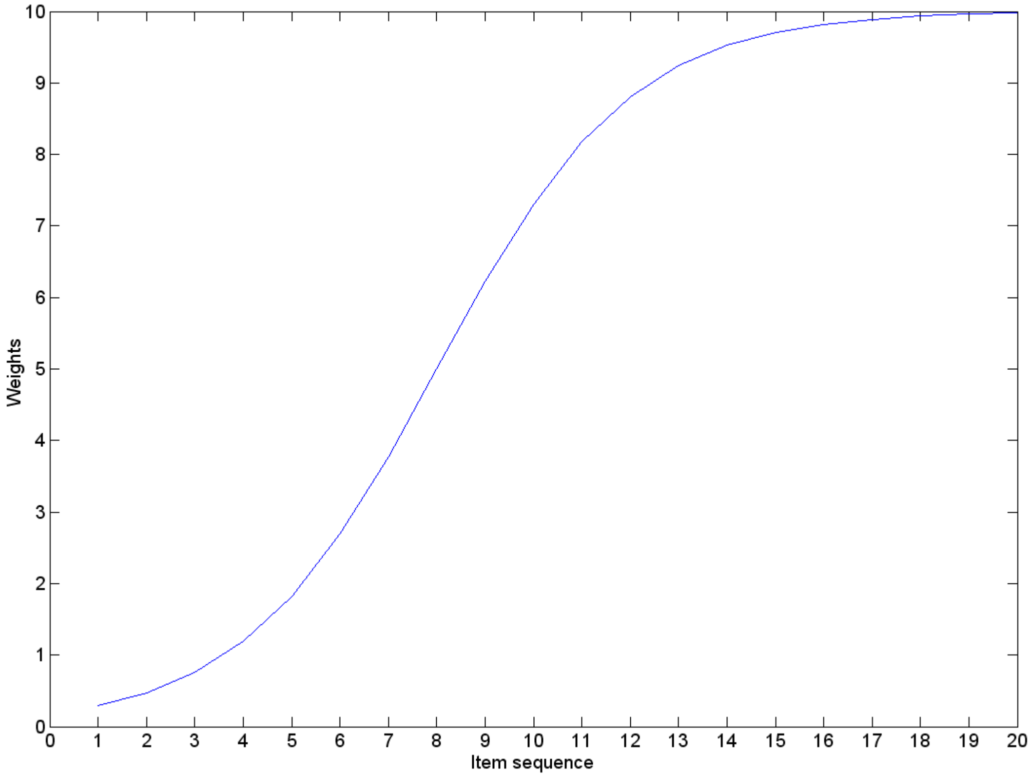


Figure 3. Proportions of tests with different numbers of constraint violations

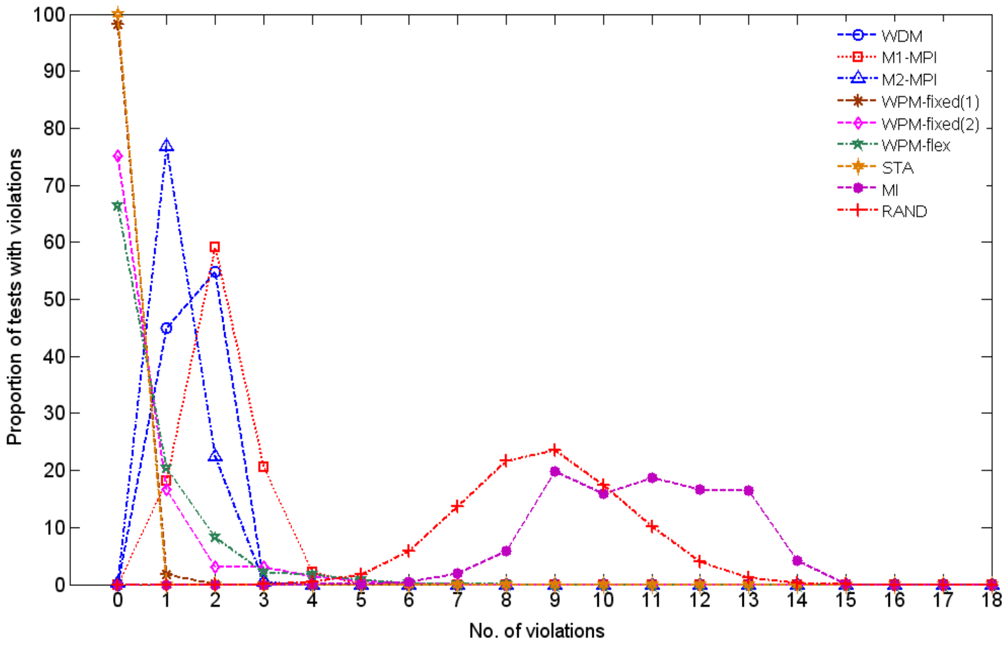


Figure 4. Conditional bias across different ability levels

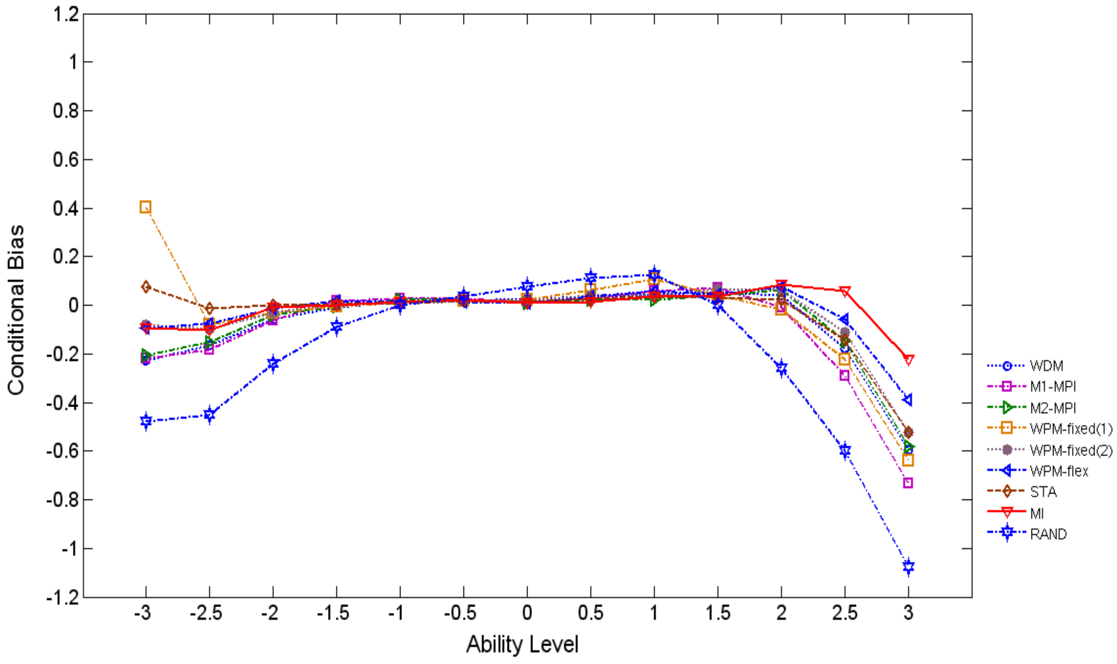


Figure 5. Conditional MSEs across different ability levels

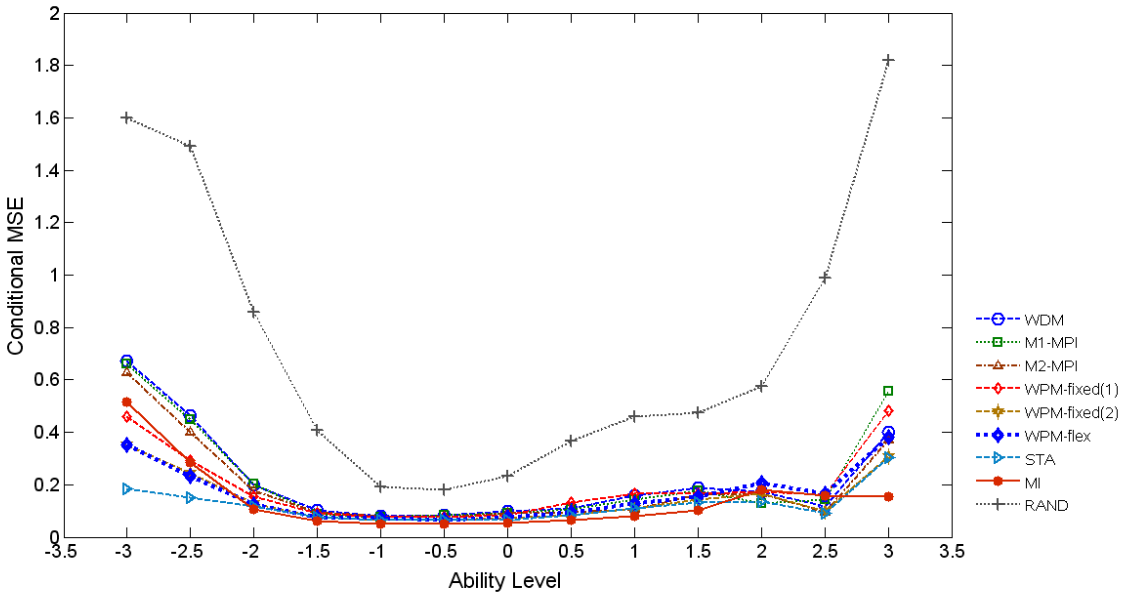
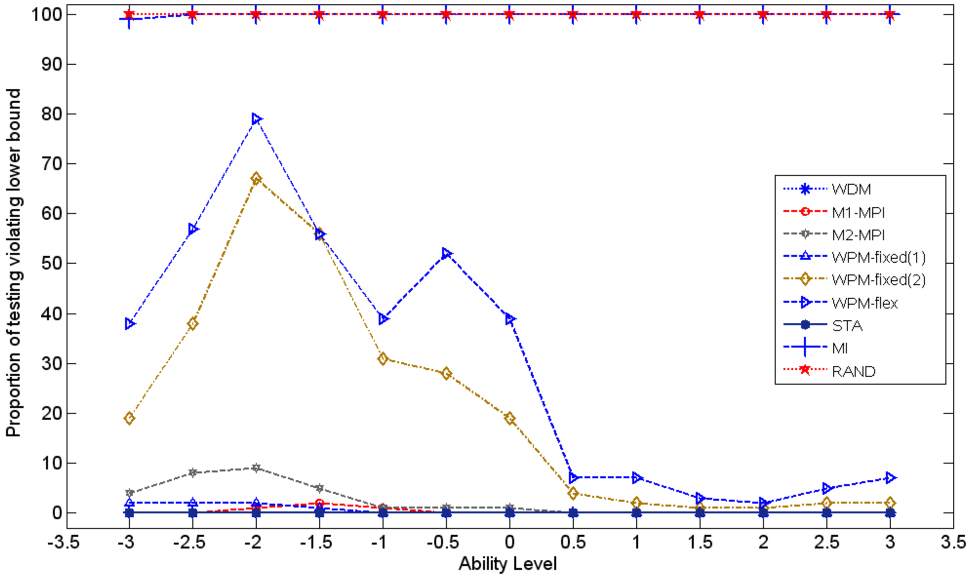
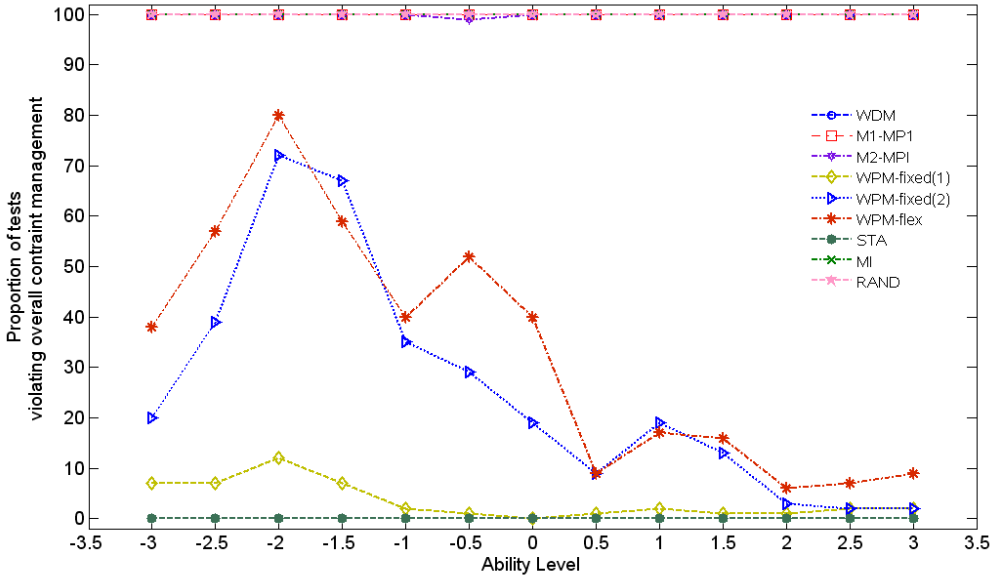


Figure 6. Proportions of tests violating lower bound



▪

Figure 7. Proportions of tests violating overall constraint management


Proportion of tests
violating overall contraint management
Ability Level
WDM
M1-MP1
M2-MPI
WPM-fixed(1)
WPM-fixed(2)
WPM-flex
STA
MI
RAND