On-the-Fly Assembled Multistage Adaptive Testing

Applied Psychological Measurement
2015, Vol. 39(2) 104–118
© The Author(s) 2014
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0146621614544519
apm.sagepub.com

Yi Zheng¹ and Hua-Hua Chang²

Abstract

Recently, multistage testing (MST) has been adopted by several important large-scale testing programs and become popular among practitioners and researchers. Stemming from the decades of history of computerized adaptive testing (CAT), the rapidly growing MST alleviates several major problems of earlier CAT applications. Nevertheless, MST is only one among all possible solutions to these problems. This article presents a new adaptive testing design, “on-the-fly assembled multistage adaptive testing” (OMST), which combines the benefits of CAT and MST and offsets their limitations. Moreover, OMST also provides some unique advantages over both CAT and MST. A simulation study was conducted to compare OMST with MST and CAT, and the results demonstrated the promising features of OMST. Finally, the “Discussion” section provides suggestions on possible future adaptive testing designs based on the OMST framework, which could provide great flexibility for adaptive tests in the digital future and open an avenue for all types of hybrid designs based on the different needs of specific tests.

Keywords

computerized adaptive testing, multistage testing, test assembly

Wainer and Eignor (2000) delineated a conjectural future where “computerized testing stations will be as ubiquitous as #2 pencils, while computing capacity and security software advance apace, p. 297.” Now their conjecture seems to have come true as the two consortia of K–12 state assessments make their transition from traditional *paper-and-pencil* (P&P) testing to computer-administered online testing. The rapidly developing computer and Internet technology can now easily support computer-administered tests designed for all kinds of purposes, from small-scale, low-stakes diagnostic tests to large-scale, high-stakes qualification tests.

Currently, there are three major designs of computer-administered tests: *computer-based tests* (CBT), *computerized adaptive tests* (CAT), and *multistage tests* (MST). CBT refers to computer-administered *linear* tests with no adaptation algorithm, whereas CAT and MST are both *adaptive*. The main goal of adaptive tests is to tailor the test to each examinee’s ability level, making it neither too easy nor too hard for him or her (Lord, 1970). Because items that are too easy or too hard provide little information for estimating the examinee’s ability level, adaptive tests can effectively reduce the test length without sacrificing much measurement

¹Arizona State University, USA

²University of Illinois at Urbana–Champaign, USA

Corresponding Author:

Yi Zheng, Arizona State University, Spark Sun Devil P.O. Box 871811, Tempe, AZ 85287, USA.

Email: yi.isabel.zheng@asu.edu

Figure 1. An multistage testing (MST) panel under 1-3-3 design.

accuracy. The proportion of test length reduction in CAT compared with linear tests of the same accuracy was found to be 30% to 50% from empirical studies (Wainer & Eignor, 2000).

CAT has been implemented in operational settings for several decades (e.g., Sands, Waters, & McBride, 1997), while MST was recently adopted by several major testing programs (e.g., the Certified Public Accountant Exam and the Graduate Record Exams) as an alternative adaptive test design. In CAT, every item is selected *on the fly* from an item bank based on the examinee's responses to previous items as well as a set of constraints such as content coverage and item exposure. In MST, items are pre-assembled into units termed *panels* (Figure 1) before administration. A panel is comprised of several stages; each stage has a certain number of *modules* (i.e., item blocks) anchored at varied difficulty levels. During the administration, each examinee is randomly assigned one of the pre-assembled parallel panels and starts with the module in Stage 1. After completing one stage, the examinee is routed to the module at the most appropriate difficulty level in the next stage. The set of modules an examinee receives in the entire test is called a *pathway* (e.g., Yan, von Davier, & Lewis, 2014).

This article presents a new adaptive testing design called *on-the-fly assembled multistage adaptive testing* (OMST), which combines the benefits of CAT and MST and offsets their limitations (Chang, 2014; Zheng & Chang, 2014; Zheng, Wang, Culbertson, & Chang, 2014). The remainder of the article describes the prototype of OMST, explains the benefits of OMST in comparison with CAT and MST, presents a simulation study that compares OMST with MST and CAT, and provides a discussion on the possible future of adaptive tests. In the future of computer- and Internet-based testing, OMST could offer great flexibility for adaptive tests and could open an avenue for all types of hybrid designs based on the different needs of the test.

OMST

Basic Framework of OMST

Like MST, OMST is administered in stages and only adapts between stages. But unlike MST, where the modules in every stage are all pre-assembled before administration, the stages in OMST are assembled on the fly.

Figure 2 illustrates the basic framework of OMST. First, the initial stage is assembled at a moderate difficulty level and administered to the examinee. A provisional ability estimate is computed based on the responses to the items in the first stage, and a new set of items is then assembled as the second stage to match the provisional ability estimate. After the examinee completes the second stage, the ability estimate is updated based on the responses to all of the administered items and a new stage is assembled to match the updated ability estimate. This process continues until the test is terminated. After the entire test is completed, the examinee is

Figure 2. A basic design of on-the-fly assembled multistage adaptive testing (OMST).

scored based on his/her responses to all of the administered items. Note that whenever a stage is assembled, computer algorithms can be used to control the content coverage, exposure rates, and other properties of the selected items. Also, although Figure 2 only illustrates three stages, OMST can have as many stages as desired (as indicated by the ellipsis on the right side of Figure 2). The numbers of items in each stage are also flexible, and there can be stopping rules to terminate the test once a satisfactory measurement accuracy is reached.

Item Selection in OMST

Tests without nonstatistical constraints. If the test does not have non-statistical constraints (e.g., content coverage, word count), any well-established item selection method in CAT can be readily adapted and used in OMST. The *maximum Fisher information* method, for example, can be used: In each stage, select a set of items one by one to maximize the information at the latest provisional ability estimate and then administer them together. Other existing CAT item selection criteria can also be used in the same way. In this situation, the only difference in item selection between OMST and CAT is that CAT adapts at the item level and OMST adapts at the stage level.

Tests with nonstatistical constraints. If the test has non-statistical assembly constraints, two strategies can be combined in OMST to assemble the stages that match the provisional ability estimate and simultaneously meet all constraints.

Strategy 1: Constraint-controlled item selection methods. Various existing constraint-controlled item selection methods developed for CAT can be adapted to OMST: To assemble a stage in OMST, simply select a set of items sequentially to optimize the objective function specified by the constraint-controlled item selection method and then administer them together.

Among the existing constraint-controlled item selection methods are two widely used approaches: the 0-1 programming approach and the heuristic approach. The 0-1 programming

approach always guarantees that all constraints are satisfied, but it is computationally intense, relies on powerful solver software, and may not have a solution when the given item bank is not sufficient for the constraints (i.e., the *infeasibility problem*). Although the 0-1 programming approach was originally introduced to assemble fixed test forms, the *shadow test* method (van der Linden, 2005, 2010) is an application of 0-1 programming into CAT item selection and can be adopted in OMST. Unlike 0-1 programming, heuristic methods do not guarantee that all constraints can be met, but they are usually much faster, do not rely on solver software, and do not have the infeasibility problem. Some useful heuristic methods for constraint-controlled item selection include the *weighted deviation model* (WDM; Swanson & Stocking, 1993), the *normalized weighted absolute deviation heuristic* (NWADH; Luecht, 1998), and the *maximum priority index* (MPI; Cheng & Chang, 2009).

Strategy 2: The item replacement step. The previously mentioned constraint-controlled item selection methods should produce stages with little violation of the constraints, if any. This may be already satisfactory to some test developers. However, if the test developer requires that all constraints must be strictly met and the chosen item selection method itself cannot guarantee that, like the heuristic methods, an item replacement step can be added after the initial item selection.

Specifically, after a set of items is selected for the next stage by the chosen constraint-controlled item selection method, these items are checked against the constraints. If any constraint is not met, some items in the current set with undesirable attributes will be replaced with desirable items from the item bank so that the constraint is met. Among the items in the item bank with desirable attributes, the replacement item can be chosen based on an information criterion to help optimize measurement accuracy. After all violations are fixed, the set of items is then administered to the examinee as the next stage.

The algorithm of the item replacement step depends on the specific types of constraints. Below is the algorithm for a simple case, all constraints being discrete lower bounds for content coverage, as an illustration.

After a set of items is selected for the next stage by the chosen constraint-controlled item selection method, compute the deviation in the content coverage of the current item set from the constraints.

WHILE ANY content category has a deficiency

Randomly select one of the categories with deficiency, category k.

Assign the number of items deficient from the lower bound in category k to n.

Select n items from the available items in the item bank that belong to category k and maximize the Fisher information at the current ability estimate.

FOR i IN the n selected replacement items

FOR j IN the randomly permuted current set of selected items

*IF item j does not belong to any of the categories with deficiency
AND the replacement of item j by item i does not result in
the deficiency of other categories,*

*THEN replace item j in the current set of selected items by item i,
and BREAK.*

END IF

END FOR

*Re-compute the deviation of the updated set of selected items from the
constraints.*

END FOR

END WHILE

Note that in CAT, the ease of computation (the advantage of the heuristic approach) and the guarantee of satisfaction of all constraints (the advantage of the 0-1 programming approach) usually cannot be achieved simultaneously. However, OMST provides a unique opportunity to achieve both: The heuristic methods can be used to give a fast and mostly satisfactory initial solution; the item replacement step can be used afterwards to guarantee all constraints are met.

Exposure Control in OMST

Item exposure can be controlled by automated algorithms in OMST like in CAT. On one hand, we should prevent the popular items from being over-exposed in case of disclosure and sharing among examinees. To serve this purpose, the *Sympson–Hetter* (SH; Sympson & Hetter, 1985) method or Stocking and Lewis's (1995) *multinomial Sympson–Hetter* (MSH) method, for example, can be used. For details of the two exposure control methods, see the original articles (Stocking & Lewis, 1995; Sympson & Hetter, 1985). The SH method is usually more effective than the MSH method in controlling the maximum exposure rate; however, when paired with the 0-1 programming approach, the SH method can significantly prolong the computation time (Veldkamp & van der Linden, 2008). Therefore, the MSH method is recommended when the 0-1 programming approach is used for item selection, and the SH method is recommended when the heuristic methods are used for item selection.

On the other hand, to reduce the number of items rarely administered, an item bank stratification method is recommended. The steps of item bank stratification are described below.

Step 1: A simulation experiment of the SH or MSH controlled OMST without item bank stratification is carried out.

Step 2: The items in the item bank are partitioned into two sub-banks—the *under-used sub-bank* and the *well-used sub-bank*—according to their initial exposure rates and their non-statistical attributes. The sizes of the two sub-banks are proportional to the number of items in the first stage versus that of all subsequent stages. To obtain the under-used bank with both the lowest possible exposure rates and a balanced composition of non-statistical attributes, a 0-1 programming problem is set up as follows:

Minimize

$$\sum_{i=1}^I p_i x_i \quad (1)$$

subject to

$$x_i \in \{0, 1\}, i = 1, \dots, I, \quad (2)$$

$$\sum_{i=1}^I x_i = \lfloor rI \rfloor, \quad (3)$$

$$\sum_{i \in V_{c_k}} x_i \geq \left\lfloor r \sum_{i \in V_{c_k}} 1 \right\rfloor, k = 1, \dots, K, \quad (4)$$

where p_i denotes the initial exposure rate of item i from the simulation experiment in Step 1; $x_i = 1$ indicates item i is included in the underused sub-bank, $x_i = 0$ otherwise; I is the total

Figure 3. Comparison of a -parameter distributions of the complete item bank and the underused item banks in CAT and OMST.

Note. CAT = computerized adaptive testing; OMST = on-the-fly assembled multistage adaptive testing.

number of items in the item bank; r is the ratio of the number of items in the first stage to the total test length; the sign “ $\lfloor \cdot \rfloor$ ” denotes the floor function, which gives the greatest integer smaller than the contained value; and V_{c_k} is the set of items relevant to constraint k . The solution to this 0-1 programming problem gives the underused bank, and the remaining items become the well-used bank.

Step 3: The underused bank is used in the first stage of OMST with random item selection, and the well-used bank is used in all subsequent stages with the chosen adaptive item selection method.

The item bank stratification procedure is developed based on the rationale of Chang and Ying’s (1999) a -stratification method. This rationale is also proved by the simulation study. According to the simulation, the only significant difference between the under-used bank and the complete bank was found in the distribution of a -parameters. Figure 3 shows the difference: The under-used bank in CAT and OMST both have relatively low a -parameters compared to the complete bank. According to Chang and Ying’s (1999) article, at the initial stage, when we have no knowledge of the examinee at all, we need low- a items to shed light on a wider range of possible ability values like a floodlight (Davey & Nering, 2002). This also saves the high- a items for later stages, when we have a naive estimation of the ability level and need high- a items to provide greater discriminating power in the neighborhood of the estimated ability location, like a spotlight (Davey & Nering, 2002).

Theoretical Comparison of OMST With CAT and MST

The development of the OMST framework was inspired by an analysis of the history of CAT and MST. As mentioned earlier, CAT has been implemented in operational settings for several decades, while MST was recently adopted as an alternative. The introduction of MST effectively alleviated some problems of the earlier applications of CAT.

Among all unexpected consequences of earlier applications of CAT, one that perhaps led to great urge for MST was the incidents in 2000 and 2003, where the GRE and the Graduate Management Admission Test (GMAT) CAT systems were found to have produced unreliable (i.e., severely underestimated or overestimated) scores for thousands of test takers (Carlson,

2000; Chang & Ying, 2008; Merritt, 2003). According to Chang and Ying (2008), the under-/overestimation problem was mainly caused by the item selection algorithm that heavily relied on maximizing the Fisher Information. Under this item selection scheme, the step size of updating ability estimates is much larger when the number of administered items is small than when it is large; therefore, if an able examinee accidentally answers the first few items incorrectly, or a less able examinee happens to guess the first few items correctly, it is difficult for the estimates to return to their true ability levels within a short test. MST is immune to this problem because it does not estimate examinee abilities until the entire first stage is completed, and therefore the step size will never be too large for the remaining test to recover an inaccurate estimate. Like MST, OMST can alleviate the over-/underestimation problem of earlier CAT in the same way.

A second major limitation of earlier CAT, according to Hendrickson (2007), is not allowing examinees to skip items or revisit completed items to prevent manipulation of the adaptive algorithms for illegitimate purposes (e.g., Kingsbury, 1996; Wainer, 1993). This usually adds unnecessary stress to examinees, which however, can be alleviated by both MST and OMST. Because MST and OMST only adapt between stages, they allow examinees to navigate back and forth freely within their current stage to review and change their answers.

Note that another well-liked feature of MST is that the test developers can review the assembled test forms and make sure all assembly requirements are met before administration. However, when there are a great number of test forms, human review of all test forms may become too cumbersome and expensive. In OMST, all these quality control tasks can be performed by computers much faster and with a lot more ease through the constrained item selection, the item replacement step, and the exposure control algorithms. As mentioned earlier, OMST provides a unique opportunity to achieve both the ease of computation and the guaranteed satisfaction of all constraints simultaneously, which is another advantage over CAT.

OMST not only shares the advantages of MST but also overcomes some limitations of MST. For example, because the stages of OMST are assembled on the fly to match each individual's ability level, OMST shall provide greater information for each individual examinee than MST given good model fit. Consequently, the measurement efficiency of OMST will be improved from MST, especially at the two polar ends of the ability scale.

Another limitation of MST is that, because items are bundled together in modules and modules are bundled together in panels, MST may be subject to a new threat to test security (Wang, Zheng, & Chang, 2014). If an MST is administered continuously and the test items are shared among friends or disclosed on the Internet, it is possible that examinees who happen to draw the same panel and take the same pathway as the discloser will be able to get almost all of the items correct regardless of their true ability levels (Wang et al., 2014). Unlike MST, OMST selects the items on the fly, so it is less likely to see two examinees share the same test in OMST than in MST.

Moreover, the assembly of multiple parallel panels for large-scale MST is a highly complex task, because the assembly procedure needs to simultaneously meet three demanding goals: (a) to make the information curves or difficulties of the *modules* in a stage *distinct enough* to provide valid adaptation, (b) to make the information curves of corresponding *pathways* across panels *similar enough* to achieve parallel panels, and (c) to meet all non-statistical constraints for every pathway in every panel (Zheng et al., 2014). In contrast, the on-the-fly assembly of OMST is more straightforward and easily implemented automatically by computer algorithms.

In summary, on one hand, OMST shares the benefits of MST including no severe over-/underestimation of examinee ability levels and allowing examinees to navigate freely within a stage; on the other hand, it also shares the benefits of CAT such as more accurate adaptation,

Table 1. Distributions of Item Parameters in the Item Bank.

<i>a</i>		<i>b</i>		<i>c</i>	
<i>Mean</i>	<i>SD</i>	<i>Mean</i>	<i>SD</i>	<i>Mean</i>	<i>SD</i>
0.754	0.404	0.049	1.278	0.123	0.048

items not bundled together, and straightforward test assembly. Furthermore, OMST has another strength that all nonstatistical constraints can be satisfied automatically with ease.

Empirical Comparison of OMST With CAT and MST: A Simulation Study

A simulation study was conducted to compare OMST with MST and CAT. There are three nested factors under OMST and CAT: (a) constrained item selection approach (0-1 programming selection vs. heuristic selection), (b) exposure control algorithm (SH vs. MSH), and (c) item bank stratification (stratified vs. non-stratified).

General Specifications

The simulation programs were written in MATLAB. The item bank consists of 352 multiple-choice items retired from a large-scale computerized adaptive English language proficiency test. The items belong to eight content categories. Each item belongs to one or more categories. All of the items are calibrated using the *three-parameter logistic* (3PL) model. The distributions of item parameters and numbers of items in each content category are summarized in Tables 1 and 2. The ability parameter values (denoted by θ) of 500 examinees were simulated from the standard normal distribution, truncated within $(-3.5, 3.5)$. The reason for truncation is to prevent confounding effects brought in by potential outliers. Fifty replications were simulated for each test design. For better comparability, in each replication, the same simulated examinees were used across all test designs.

The test length was fixed at 45 items. For the tests administered in stages (i.e., OMST and MST), the number of stages was set to three and each stage had 15 items. There is one non-statistical constraint based on the content categories: For CAT, the entire test was required to have at least three items from each category; for OMST and MST, each stage was required to have at least one item from each category.

The examinee's ability, θ , was estimated by a combination of *maximum likelihood estimation* (MLE) and *expected a posteriori* (EAP) estimation. MLE is the best unbiased estimator for θ but is not available when the responses from an examinee are all correct or all incorrect. So when the responses are all correct or incorrect, EAP was used to estimate θ ; otherwise, MLE was used.

MST Specifications

The MST panel design used in the comparison study is the 1-3-3 design as shown in Figure 1. The difficulty of the module in the first stage was anchored at 0 on the ability scale. The difficulties of the modules in the second and third stages were anchored at -1 , 0 , and 1 .

Table 2. Number of Items in Each Content Category in the Item Bank.

Category	1	2	3	4	5	6	7	8
No. of items	57	46	80	60	40	20	40	32

The MST panels were assembled using the *bottom-up* approach (Luecht & Nungester, 1998). First, multiple parallel modules of each module position were assembled from the item bank; then the parallel modules were mixed-and-matched to construct multiple parallel panels. To obtain a balanced item exposure and to use as many items from the item bank as possible, 11 parallel forms were assembled for the module in the first stage and 2 parallel forms were assembled for each module in the second and third stages (Zheng, Nozawa, Gao, & Chang, 2012). The resulting total number of active items in MST is 345. With the bottom-up approach, theoretically a large number of parallel panels can be assembled at maximum (704 in this study), but in a practical sense, only a certain number of panels are assembled and used. In this study, 50 parallel panels were randomly selected from all of the possible panels. The modules were assembled using a simplified version of the NWADH (Luecht, 1998). The detailed algorithm for assembling MST in this simulation study is given in Online Appendix A. Interested readers can also refer to Zheng et al. (2012).

OMST and CAT Specifications

The simulation study evaluated both the heuristic and 0-1 programming approaches as the constraint-controlled item selection method in OMST and CAT. For the heuristic approach, the MPI method (Cheng & Chang, 2009) was used as a representative. For the 0-1 programming approach, the shadow test method (van der Linden, 2005, 2010) was used. The detailed algorithms of the MPI method and the shadow test method as implemented in the simulation study are given in Online Appendices B and C. For all OMST conditions, the item replacement step as mentioned earlier in this article was added before the administration of every stage to ensure that all constraints are met.

To prevent popular items from being overexposed, either the SH method (Simpson & Hetter, 1985) or the MSH method (Stocking & Lewis, 1995) was imposed for OMST and CAT. (The item exposure control in MST is carried out in the test assembly stage.) As mentioned earlier, when the 0-1 programming approach is used for item selection, the SH method can significantly prolong the computation time, so only the MSH method was paired with the shadow test item selection. Both the SH and MSH were paired with the MPI item selection because SH is more effective and MSH provides a fairer comparison with the shadow test item selection. The threshold parameter (r) was set to 0.15 for both SH and MSH, and the calibration phase had 10 iterations. The size of the “list of items” in the MSH method was set to eight, which is the ratio between the item bank size and the test length as suggested by Stocking and Lewis (1998). The MSH algorithm was also adapted when the shadow test method was used: If the number of free items in the shadow test is less than eight but greater than one, the free items form the “list of items.” If there is only one free item in the shadow test, the SH method is used instead of the MSH method.

Both CAT and OMST were implemented with or without item bank stratification. In the stratified conditions, the first stage of the OMST and first 15 items of the CAT (equivalent length with the first stage of the OMST) were selected from the underused sub-bank, and the remaining 30 items were selected from the well-used sub-bank as described earlier.

Results

The results of the 50 replications are averaged and summarized in Tables 3 and 4. The following sub-sections will discuss the results from different perspectives.

Measurement Accuracy and Maximum Exposure Rate

The measurement accuracy of the tests is evaluated by the root mean square error (RMSE) and bias of examinee ability θ estimates against the true θ values:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{\theta}_i - \theta_i)^2}, \quad (5)$$

$$\text{Bias} = \frac{1}{n} \sum_{i=1}^n (\hat{\theta}_i - \theta_i), \quad (6)$$

where n is the number of examinees, θ is the true value of the examinee ability, and $\hat{\theta}$ is the estimate of examinee ability. The smaller the RMSE and the absolute values of bias are, the better the measurement accuracy is.

The item exposure rate is computed for each item by dividing the number of examinees who answered this item by the total number of examinees. The maximum item exposure rate reflects how well the algorithm controls the overexposure of popular items.

Table 3 contains the RMSE and bias values and maximum item exposure rates of the 13 compared tests. In general, there is a trade-off between the maximum item exposure rate and RMSE: If the maximum item exposure rate of a test is controlled to be smaller, the test will have a larger RMSE; in other words, measurement accuracy will be sacrificed to achieve better test security. This trade-off is seen in most pairs of test designs in the simulation study. There are a few exceptions, though. An interesting case is OMST versus CAT with non-stratified MPI item selection and SH exposure control. In this case, OMST slightly outperforms CAT in both RMSE and maximum item exposure rate. This is somewhat counterintuitive given a perfect model fit and zero nonstatistical constraint. However, the real practice is always subject to noises and constraints, and therefore, the simulation results have suggested the promise of OMST in practical setting.

We have two other observations. First, all conditions of OMST and CAT have similar amounts of bias, while MST produced more bias than OMST and CAT. Second, the SH exposure control method was effective in controlling the maximum item exposure rates under a reasonable threshold, whereas the MSH method was less effective. As mentioned earlier, the reason for using the MSH exposure control is that the SH method can significantly prolong the computation time of the shadow test item selection (Veldkamp & van der Linden, 2008). But if any heuristic method is chosen as the item selection method instead of a 0-1 programming method, then the regular SH exposure control is recommended.

Constraint Violations

We also evaluated how well the item selection algorithms met the nonstatistical constraints by (a) the constraint violation rate, which is the average occurrence rate of the violation of a content constraint, and (b) the violation-free rate, which is the proportion of tests with zero violation of all constraints. In Table 3, we can see that all OMST conditions generated violation-free tests. This is guaranteed by the item replacement step uniquely made possible by OMST. This

Table 3. Evaluation Indices of the Compared Test Designs (Part I).

	RMSE	Bias	Violation rate	Violation-free rate	Maximum exposure rate	Item bank size	No. of items unused
MST	0.498	0.063	0.063	0.495	0.205	345	0.0
OMST							
Non-stratified							
MPIR-SH	0.479	0.023	0	1.000	0.266	352	41.5
MPIR-MSH	0.333	0.020	0	1.000	0.976	352	189.9
STR-MSH	0.396	0.021	0	1.000	0.630	352	122.9
Stratified							
MPIR-SH	0.481	0.028	0	1.000	0.215	352	10.8
MPIR-MSH	0.378	0.024	0	1.000	0.563	352	81.3
STR-MSH	0.430	0.026	0	1.000	0.409	352	43.0
CAT							
Non-stratified							
MPI-SH	0.483	0.021	0.012	0.905	0.278	352	34.6
MPI-MSH	0.321	0.018	0	1.000	0.817	352	172.9
ST-MSH	0.400	0.021	0.000 ^a	1.000 ^a	0.371	352	104.6
Stratified							
MPI-SH	0.491	0.023	0.012	0.907	0.220	352	0.7
MPI-MSH	0.345	0.023	0	1.000	0.639	352	72.2
ST-MSH	0.417	0.028	0	1.000	0.393	352	22.7

Note. RMSE = root mean square error; MST = multistage testing; OMST = on-the-fly assembled multistage adaptive testing; MPIR = maximum priority index item selection (MPI) with the item replacement step; SH = the regular Simpson–Hetter exposure control method; MSH = the multinomial Simpson–Hetter exposure control method; STR = shadow test item selection (ST) with the item replacement step; CAT = computerized adaptive testing; MPI = maximum priority index item selection; ST = shadow test item selection.

^aThe violation rate value here is 1.00E-05, and the violation-free rate value here is 0.99996. Note that normally the shadow test method will guarantee zero violation of all constraints. The reason for the slight violation here is that when the shadow test is combined with the SH exposure control, if none of the available items pass the probability experiment in the SH method, the program will allow a onetime violation.

Table 4. Evaluation Indices of the Compared Test Designs (Part II).

	Overall overlap rates		Conditional overlap rates					
			Mean			SD		
	Mean	SD	Low	Medium	High	Low	Medium	High
MST	0.144	0.217	0.230	0.174	0.226	0.255	0.224	0.253
OMST								
Non-stratified								
MPIR-SH	0.155	0.054	0.164	0.158	0.164	0.055	0.054	0.057
MPIR-MSH	0.506	0.162	0.618	0.633	0.595	0.113	0.103	0.108
STR-MSH	0.245	0.082	0.297	0.274	0.298	0.075	0.065	0.087
Stratified								
MPIR-SH	0.143	0.060	0.165	0.148	0.173	0.060	0.053	0.072
MPIR-MSH	0.260	0.144	0.334	0.323	0.325	0.125	0.115	0.133
STR-MSH	0.174	0.084	0.212	0.196	0.210	0.082	0.069	0.095
CAT								
Non-stratified								
MPI-SH	0.151	0.063	0.161	0.154	0.161	0.064	0.063	0.068
MPI-MSH	0.453	0.238	0.623	0.656	0.580	0.153	0.124	0.166
ST-MSH	0.219	0.088	0.273	0.255	0.268	0.075	0.063	0.099
Stratified								
MPI-SH	0.142	0.055	0.154	0.144	0.156	0.057	0.053	0.059
MPI-MSH	0.324	0.188	0.462	0.465	0.434	0.133	0.112	0.145
ST-MSH	0.178	0.084	0.237	0.213	0.219	0.081	0.059	0.095

Note. MST = multistage testing; OMST = on-the-fly assembled multistage adaptive testing; MPIR = maximum priority index item selection (MPI) with the item replacement step; SH = the regular Simpson–Hetter exposure control method; MSH = the multinomial Simpson–Hetter exposure control method; STR = shadow test item selection (ST) with the item replacement step; CAT = computerized adaptive testing; MPI = maximum priority index item selection; ST = shadow test item selection.

item replacement step cannot be applied in CAT because CAT administers every item once they are selected. Thus, zero violation is not always guaranteed in CAT. Note that the violation rates are all considerably small due to the strength of the constraint-controlled item selection methods. However, if the test developer requires that all constraints must be strictly met, OMST can easily satisfy this requirement. More violation is reported for MST. In MST practice, these violations can be fixed by human review of the assembled panels.

Item Bank Stratification

The item bank stratification strategy was designed to improve the utilization of the less popular items in the item bank. This is evaluated through the number of unused items. Comparing the tests with and without item bank stratification, results in Table 3 show that the stratification design greatly reduced the number of unused items while maintaining a similar level of measurement accuracy in both OMST and CAT. This resonates with the original study by Chang and Ying (1999).

Test Security

Test security is evaluated through the means and standard deviations of the *pairwise test overlap rates*. The pairwise test overlap rate is computed as the proportion of overlapping items between any possible pair of examinees. The mean and the standard deviation over all possible pairs are

reported as the overall test overlap. Three subgroups are also formed for computing conditional test overlap rates based on the true θ values: the lowest about 33% (i.e., $\theta < -0.43$), the middle about 33% (i.e., $-0.43 \leq \theta < 0.43$), and the highest about 33% (i.e., $\theta \geq 0.43$).

Traditionally, only the mean of pairwise test overlap rate is used to evaluate test security. As shown in Tables 3 and 4, the means of the overall test overlap rates are closely related to the maximum item exposure rate. The lower the maximum item exposure rate is, the more even the usage of the item bank is, and the smaller the mean of test overlap rates is. This relation is expected and has been mathematically proved by Chen, Ankenmann, and Spray (2003).

However, a more important finding lies in the standard deviations of pairwise test overlap rate. Wang et al. (2014) proved that the standard deviation of pairwise test overlap rate provides important supplementary information for test security. Even if the means of test overlap rates are the same, different magnitudes of standard deviations could reveal different security statuses. This is especially important when we compare MST with the on-the-fly adaptive tests (e.g., OMST and CAT), because MST bundles items together in modules, which will produce larger standard deviations of test overlap rates than the on-the-fly adaptive tests with the same mean of test overlap rates. If the test items are disclosed, a large standard deviation reflects that some examinees may get a huge benefit from the disclosed information.

This is empirically verified by the results shown in Table 4. Take the OMST and CAT conditions with MPI item selection and SH exposure control as a comparison with MST. Their average test overlap rates are on similar levels. If judging from the average test overlap rates only, one would conclude that they are similarly secure. However, a salient difference appears in the standard deviations of test overlap rates: The standard deviations of test overlap rates in MST are about 4 times of those in the compared conditions in OMST and CAT. In MST, a considerable number of examinee pairs are administered with highly overlapped tests, the maximum overlap rate being 100% (same pathway in the same panel), whereas CAT and OMST both produce highly individualized, diverse test forms. These results empirically proved the existence of the new security concern brought in by the feature of item bundling in MST and that both OMST and CAT are immune to this problem.

Discussion

This article proposed a new adaptive test design called “on-the-fly multistage adaptive testing” (OMST), which could combine the merits of CAT and MST and offset their limitations. A prototype design of OMST was presented and demonstrated through a simulation study. The simulation results show that (a) the measurement accuracy of OMST is comparable to that of CAT and MST; (b) the constraint-controlled item selection method and the item replacement step together can generate constraint violation-free OMST; (c) the item bank stratification method was effective in increasing the usage of less popular items without sacrificing too much measurement accuracy; (d) the SH item exposure control was able to keep the maximum item exposure rates at a desirable level, while the MSH method was less effective; and (e) the tests assembled by OMST and CAT are highly diverse, leading to potentially better test security than MST.

Note that the current article only presents a basic design of OMST, and this design can be further varied or improved in many aspects. One possible variation is variable-length OMST. One major advantage of CAT is that it shortens the test without sacrificing measurement accuracy, and variable-length CAT can further strengthen this advantage by terminating the test even earlier for certain examinees once satisfactory measurement accuracy is achieved. This is also applicable to OMST, and hence, the development of termination methods is needed. Intuitively, the existing termination criteria for CAT can also be applied in OMST in a fairly straightforward manner. The only difference is that OMST terminates after the examinee completes all the

items within a stage, whereas CAT can terminate after the examinee completes any single item if rules allow.

Another possible generalization is that the stage length in OMST can vary by stages. One plausible design is to gradually shrink the stage length as the test proceeds. In the first stage, when no information is available about the examinee's ability level, a longer stage can be useful in providing a better ability estimate for the adaptive item assembly. In later stages, as the estimate gets closer to its true value, shorter stage lengths will lead to more chances to adapt, and consequently higher efficiency. This can be regarded as a smooth transition from MST to CAT or a "hybrid design."

The number of stages in OMST can also vary. The simulation study in this article only evaluated the three-stage design because the main purpose of the article is to present a prototype of OMST. However, it is expected that the more stages an OMST has, the closer its properties resemble a CAT, and the fewer stages the OMST has, the closer its properties resemble an MST.

Note that albeit the ostensible difference between CAT and MST, they were in fact both developed from *sequential analysis* (Wald, 1945) and closely related. CAT can be regarded as "individual sequential," while MST as "group sequential." In this light, OMST can also be regarded as group sequential, and the above-mentioned "hybrid design" could be a promising direction for future adaptive testing design, which unifies CAT and MST and provides the most flexibility to various test development requirements.

Finally, the results of the three testing modes are all highly dependent on the quality of the given item bank and the specific designs. Although this study tried to present a fair comparison, potential improvements or changes in test designs, such as test assembly methods and routing rules in MST, can certainly vary the results. Therefore, the results of this study by no means suggest that one testing mode is superior to the others; rather, this article is purported to propose a new option for computer-assisted adaptive testing and hopefully lead to more innovations and improvements.

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

- Carlson, S. (2000). ETS finds flaws in the way online GRE rates some students. *Chronicle of Higher Education*, 47(8), A47.
- Chang, H.-H. (2014). Psychometrics behind computerized adaptive testing. *Psychometrika*. Advance online publication. doi: 10.1007/s11336-014-9401-5
- Chang, H.-H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement*, 20, 213-229.
- Chang, H.-H., & Ying, Z. (1999). *a*-stratified multistage computer adaptive testing. *Applied Psychological Measurement*, 23, 211-222.
- Chang, H.-H., & Ying, Z. (2008). To weight or not to weight? Balancing influence of initial items in adaptive testing. *Psychometrika*, 73, 441-450.
- Chen, S. Y., Ankenmann, R. D., & Spray, J. A. (2003). The relationship between item exposure and test overlap in computerized adaptive testing. *Journal of Educational Measurement*, 40, 129-145.

- Cheng, Y., & Chang, H.-H. (2009). The maximum priority index method for severely constrained item selection in computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology*, 62, 369-383.
- Davey, T., & Nering, N. (2002). Controlling item exposure and maintaining item security. In C. N. Mills, M. T. Potenza, J. J. Fremer, & W. C. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 1651-1691). Mahwah, NJ: Erlbaum.
- Hendrickson, A. (2007). An NCME instructional module on multistage testing. *Educational Measurement: Issues and Practice*, 26, 445-442.
- Kingsbury, G. G. (1996, April). *Item review and adaptive testing*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York, NY.
- Lord, M. F. (1970). Some test theory for tailored testing. In W. H. Holzman (Ed.), *Computer assisted instruction, testing, and guidance* (pp. 1391-1383). New York, NY: Harper & Row.
- Luecht, R. M. (1998). Computer-assisted test assembly using optimization heuristics. *Applied Psychological Measurement*, 22, 224-236.
- Luecht, R. M., & Nungester, R. J. (1998). Some practical examples of computer-adaptive sequential testing. *Journal of Educational Measurement*, 35, 229-249.
- Merrit, J. (2003). Why the folks at ETS flunked the course – a tech-savvy service will soon be giving B-school applicants their GMATs. *Business Week*, December 29, 2003.
- Sands, W. A., Waters, B. K., & McBride, J. R. (1997). *Computerized adaptive testing: From inquiry to operation*. Washington, DC: American Psychological Association.
- Stocking, M. L., & Lewis, C. (1995). *A new method of controlling item exposure in computerized adaptive testing* (ETS Research Report RR-95-25). Princeton, NJ: Educational Testing Service.
- Stocking, M. L., & Lewis, C. (1998). Controlling item exposure conditional on ability in computerized adaptive testing. *Journal of Educational and Behavioral Statistics*, 23, 57-75.
- Swanson, L., & Stocking, M. L. (1993). A model and heuristic for solving very large item selection problems. *Applied Psychological Measurement*, 17, 151-166.
- Sympton, J. B., & Hetter, R. D. (1985, October). *Controlling exposure-rates in computerized adaptive testing*. Paper presented at the annual meeting of the Military Testing Association, San Diego, CA.
- van der Linden, W. J. (2005). *Linear models of optimal test design*. New York, NY: Springer.
- van der Linden, W. J. (2010). Constrained adaptive testing with shadow tests. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 31-55). New York, NY: Springer.
- Veldkamp, B. P., & van der Linden, W. J. (2008). Implementing Sympton-Hetter item-exposure control in a shadow-test approach to constrained adaptive testing. *International Journal of Testing*, 8, 272-289.
- Wainer, H. (1993). Some practical considerations when converting a linearly administered test to an adaptive format. *Educational Measurement: Issues and Practice*, 12, 15-20.
- Wainer, H., & Eignor, D. (2000). Caveats, pitfalls, and unexpected consequences of implementing large-scale computerized testing. In H. Wainer, Dorans, N. J., R. Flaugher, B. F. Green, & R. J. Mislevy (Eds.), *Computerized adaptive testing: A primer* (pp. 171-199). Mahwah, NJ: Erlbaum.
- Wald, A. (1945). Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16, 117-186.
- Wang, C., Zheng, Y., & Chang, H.-H. (2014). Does standard deviation matter? Using “standard deviation” to quantify security of multistage testing. *Psychometrika*, 79(1), 154-174.
- Yan, D., von Davier, A. A., & Lewis, C. (2014). *Computerized multistage testing: Theory and applications*. New York, NY: CRC Press.
- Zheng, Y., & Chang, H.-H. (2014). Multistage testing, on-the-fly multistage testing, and beyond. In Y. Cheng & H.-H. Chang (Eds.), *Advancing methodologies to support both summative and formative assessments* (Chapter 2, pp. 21-39). Charlotte, NC: Information Age.
- Zheng, Y., Nozawa, Y., Gao, X., & Chang, H.-H. (2012). *Multistage adaptive testing for a large-scale classification test: The designs, automated heuristic assembly, and comparison with other testing modes* (ACT Research Reports 2012-6). Retrieved from http://media.act.org/documents/ACT_RR2012-6.pdf
- Zheng, Y., Wang, C., Culbertson, M. J., & Chang, H.-H. (2014). Overview of test assembly methods in multistage testing. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing: Theory and applications* (Chapter 6, pp. 87-99). New York, NY: CRC Press.