Research and Practice in Technology Enhanced Learning
Vol. 5, No. 1 (2010) 3–17

DOI: 10.1142/S1793206810000785

MIXED APPROACH FOR ITEM SELECTION IN E-TESTING

CHI-KEUNG LEUNG
Department of Mathematics and Information Technology
The Hong Kong Institute of Education
10 Lo Ping Road, Tai Po, Hong Kong
ckleung@ied.edu.hk

Learning and instruction can be enhanced with the information provided from assessment. Computerized adaptive testing is an effective mode of e-testing as it can be operated in a continual and dynamic mode. It can also cater for individual needs by providing tailor-made tests. In this paper, we will examine a new item selection approach which would allow more flexibility in balancing the two competitive goals of accuracy and item pool utilization in computerized adaptive testing. Moreover, a new index for comparing a weighted balance of these two goals is proposed.

Keywords: Computerized adaptive testing; catering for individual needs; item selection.

1. Introduction

Learning and instruction can be enhanced by computer technology through various means such as computer-assisted learning (CAL) and computer-assisted instruction (CAI). In fact, computer technology can bring in other advantages such as developing a continual and dynamic approach to educational testing. With the advances in computer technology and statistical modeling, more and more paper and pencil (P&P) tests have been transformed into the format of computerized adaptive testing (CAT; see, e.g. Armstrong *et al.*, 2005; Wainer *et al.*, 1990). Nowadays, various large-scale testing programs (e.g. GRE, GMAT, TOEFL and LSAT) have been partly or completely conducted in the form of CAT. One of the main advantages of CAT over P&P is that it enables more efficient and precise ability estimation (Weiss, 1982) in a continual and dynamic mode.

Currently, most of the prevalent item selection algorithms are based on the traditional wisdom of the maximum information approach, that is, the most informative item is chosen at each step of testing. The possible efficiency of the information approach, however, is likely at the expense of item security and the cost effectiveness in item pool management. In this regard, many CAT practitioners are looking for an item selection approach that can strike a balance on measurement accuracy, item security and item pool utilization. This study aimed to investigate

whether a new mixed item selection approach meets this practical need. Moreover, a new statistic has been developed in the study for evaluating overall performance of each selection method on both accuracy and item pool utilization. This statistic would facilitate practitioners to identify appropriate selection methods for their own e-testing programs.

2. Potential Advantages of Computerized Adaptive Testing

It must be emphasized that computerized adaptive testing is different from computerized administrated testing which usually refers to a mechanism that randomly select a test item or a subtest from a pool of items with regardless of the ability of the examinee (see, e.g. Beevers *et al.*, 1995). In contrast, a CAT system *adaptively selects* an item according to the estimate of the ability of examinee based on his or her responses to previous items. In other words, CAT is a dynamic system which can provide tailor-made tests for individuals.

Straetmans and Eggen (1998) broadly describe the administration procedure of a computerized adaptive testing as follows:

(1) The computer selects an item from a pool of items.
(2) The item is displayed to the examinee on the computer screen.
(3) The examinee responds to the question by typing or selecting an answer.
(4) The computer evaluates the response as correct or incorrect.
(5) If the answer is correct, the next item presented will be more difficult; otherwise an easier item will be administered.
(6) The computer terminates testing when pre-specified stopping rule is satisfied.

A few web-based testing systems were built on the theoretical bases of computerized adaptive testing (see, e.g. Conejo *et al.*, 2004; Ueno, 1998). With the support of these systems, teachers worldwide can define their tests, and their students can take these tests online adaptively.

In addition to efficiency and accuracy in assessing student achievement, Leung (1998) further elaborates that CAT has many other potential advantages such as:

(1) Students can have more flexibility to schedule their time in taking the test as computers support on-demand test delivery.
(2) Students can plan better as they will be informed the result by the computer immediately after the completion of the test.
(3) Teachers are freed from the laborious tasks of test construction and marking.
(4) Alternate item forms that involve graphics, sounds, video and text are feasible; thus some items on 3D geometry are now feasible in CAT setting.

In fact, students, teachers, school administrators and officials of education department will all benefit from a well-developed CAT system coordinated by the government or central testing organization. Particularly, student learning will be greatly enhanced. Because of the adaptive nature of CAT, students always face

items that closely match their own individual ability. At the end of the test, no one is likely to get all answers wrong and scores zero mark; the less competent students would find some items that they could solve and hence retain their interest in the learning of the subject. Neither anyone is likely to get all answers correct and scores full mark; thus even the top students understand that there are rooms for improvement. Furthermore, students spend less time on regular assessment as there are fewer inappropriate items administered by CAT for each individual.

It may happen that two test-takers get the same number of items correct, however they would have different scores as the scores depend on the calibrated item parameters like difficulty, discrimination and guessing of the items (Lord, 1980; Hambleton & Swaminathan, 1985). All the conversions of score into the same continuum are done by the statistical procedures of the system. With this ability continuum, individual's learning progress can be reported less ambiguously. It should be noted that the common impression that a student makes improvement when there is an increase in his/her scores in two successive tests is not always correct unless the two tests are comparable or linked to the same ability continuum.

Continual assessment is an integral part of teaching and learning. A well developed CAT system would be able to immediately issue individual report on one's performance. Hence, the strengths and weaknesses of the students would be identified. This timely information is very useful for making decision on instruction and student learning. If a student has unsatisfactory result in a test, follow-up actions can be determined promptly. The student can re-take the test at the time when he or she feels confident after revision or remedial teaching. Once the students are familiar with the testing procedures and environment, teachers need not accompany the students in their second and subsequent trials. Students just need to book the computers and get the teacher's approval in advance. If the item bank is rich, the test items at various attempts are very unlikely the same. With a simple indexing mechanism, an item would be administered at most once to a frequent user. Since the students themselves can determine the dates for subsequent attempts after failure, their motivation of learning would be stronger when their sense of ownership of learning builds up. Self-regulated learners tend to monitor and evaluate their learning processes (Boekaerts & Corno, 2005). With CAT, each student can be fully informed of his or her progress throughout the year. This can certainly facilitate students to become self-regulated learners.

3. Item Selection

3.1. *Maximum information approach (MI)*

Most of the CAT designs are built on the Item Response Theory (IRT; Lord, 1980). Weiss and Kingsbury (1984) described CAT as a combination of IRT, adaptive testing, and interactive computer administration of tests. When the three-parameter logistic (3-PL) IRT model is used, the probability that an examinee of ability θ

answers an item with difficulty parameter b correctly ($Y = 1$) is

$$P(Y = 1|\theta) = c + (1 - c)\frac{1}{1 + e^{-1.7a(\theta - b)}}, \tag{3.1}$$

where a and c are discrimination and pseudo-guessing parameters respectively.

Figure 1 illustrates how the probability of giving a correct answer varies with the ability, following Eq. (3.1), for four items with different set of parameters as shown in Table 1.

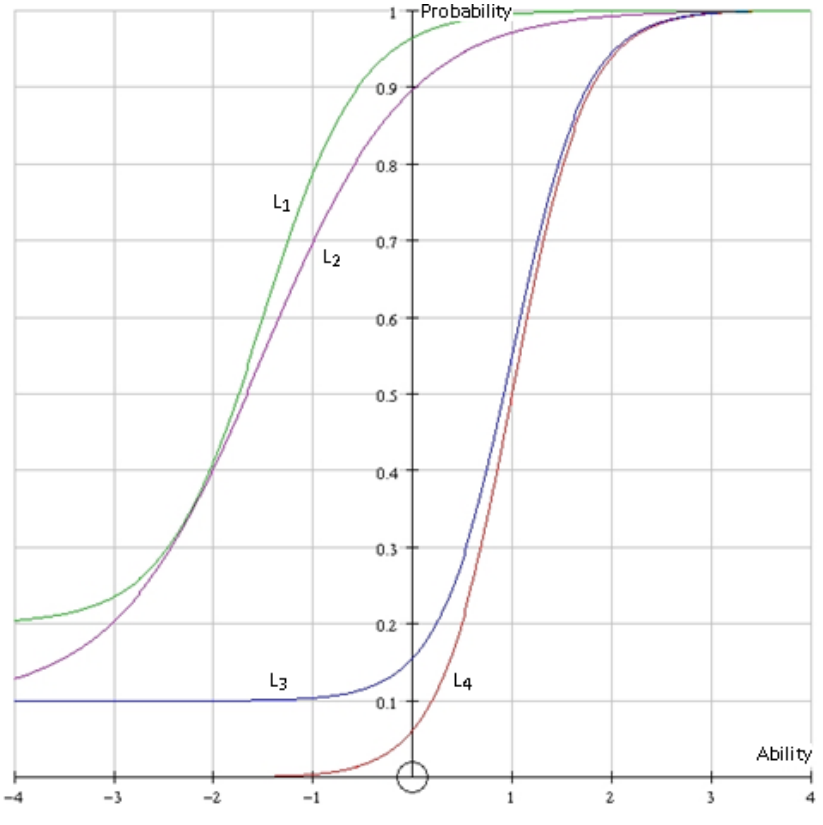

Figure 1. Variation of probability of giving correct answer against ability#.

#L_1 is for Item 1; L_2 for Item 2; L_3 for Item 3; L_4 for Item 4.

Table 1. Item parameters for the four items in Fig. 1.

Item i	a_i	b_i	c_i
1	1.2	−1.5	0.2
2	0.8	−1.5	0.1
3	1.6	1.0	0.1
4	1.6	1.0	0.0

The common features of the curves are: (a) the probability of giving correct answers lies between 0 and 1, and (b) the higher the ability, the larger the possibility of giving correct answers. Comparing to Items 3 and 4 for which the item difficulty is 1.0, Items 1 and 2 are easier as they have smaller item difficulty of -1.5. The curves for L_1 and L_2 are mainly above those of L_3 and L_4, showing that examinees of low ability have higher chance of giving correct answers to Items 1 and 2.

Item 1 can better discrimate high abilities from the low ones than Item 2 as it has larger discrimination value. The curve for L_2, compared to L_1, rises sharply when the ability exceeds the item difficulty (-1.5). Items 3 and 4 differ in having different pseudo-guessing values. The pseudo-guessing value for Item 4 is 0, meaning that apparently there is no chance of giving a correct answer to this item simply by guessing. As shown by L_4, the probability of giving correct answer approaches 0 when ability decreases. In contrast, the pseudo-guessing value for Item 3 is 0.1. It means that even an examinee has no ideas on answering the item, he or she could correctly answer it by guessing. This pattern is reflected by L_3 in which the probability of giving correct answers goes to 0.1 for examinees of low ability.

In terms of the 3-PL model, the Fisher information (also known as item information) of the item is a function of examinee's ability and is expressed as

$$I = \frac{[P'(\theta)]^2}{P(\theta)[1 - P(\theta)]}. \tag{3.2}$$

By substitution of $P(\theta)$ using Eq. (3.1) and then using differentiation, item information can be written as

$$I = \frac{1.7^2 a^2 (1 - c)}{(1 + e^{-1.7a(\theta - b)})^2 (c + e^{1.7a(\theta - b)})}. \tag{3.3}$$

Figure 2 shows the variation of information with item discrimination for a fixed value of c. The five information curves, following Eq. (3.3), correspond to cases having the same typical c of 0.2 but five different values of $(\theta - b)$ equal to $-2, -1, 0, 1$, and 2 respectively. These curves help us understand the role of a on item information.

In theory and practice, the maximum information method attempts to select an item with the largest value of a and with b closest to the examinee's ability θ. The benefit of this approach can be easily visualized in Figure 2. When item difficulty is exactly the same as the examinee's ability, i.e. $(\theta - b) = 0$, the information rises sharply as discrimination increases, following the trend of the line L_1. Thus, if the true ability θ_o is known, the maximum information approach leads to a substantial efficiency gain by choosing an item with b close to θ_o and with largest possible a. Nevertheless, θ_o can never be known in reality. So instead of θ_o, the estimated ability, $\hat{\theta}$, is used in the information calculation. During the initial stages of testing, the discrepancy between $\hat{\theta}$ and θ_o is usually large, which will gradually diminish as the testing continues. As a result, the b of the chosen item which seems to be closest to $\hat{\theta}$ may be actually far away from the true θ_o and hence the expected efficiency gain of maximum information approach cannot be realized. For example, the line

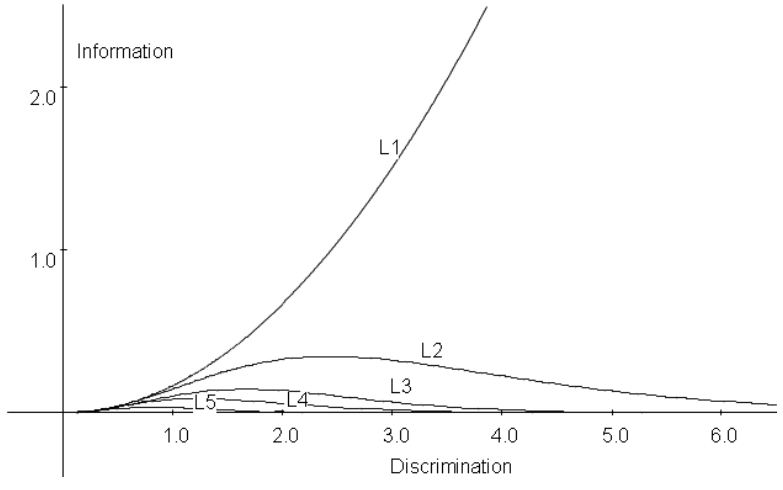

Figure 2. Information curves for five scenarios of $(\theta - b)^{\#}$.

$^{\#}c = 0.2$; L$_1$: $\theta - b = 0$, L$_2$: $\theta - b = 1$, L$_3$: $\theta - b = -1$, L$_4$: $\theta - b = 2$, L$_5$: $\theta - b = -2$

L_3 indicates that an item with a larger a of 3.0 actually provides less information than the one with a smaller a of 1.0 does when $(\theta_o - b)$ is beyond -1.0. In fact, the information goes to zero as a goes to infinity. The implication is that the maximum information approach sometimes misuses and thus over-exposes the high a items which may not be so informative as originally expected, particularly at early stages of testing when θ estimation is still quite inaccurate.

One of the major shortcomings of MI is poor item pool utilization caused by uneven usage of individual items (Stocking & Lewis, 1998; van der Linden, 1998). Under MI, many high a items are over-exposed to a high proportion of examinees, while a substantial part of the item pool is hardly used. On the one hand, over-exposed items can cause a security problem. If an item has a high exposure rate, then it has a greater risk of being known to prospective examinees, which in turn would cause item security and test validity problems. On the other hand, under-utilized items mean a waste of money as the development of these items is very costly. Therefore, MI has undesirable consequences of poor item pool utilization and item security.

3.2. *b-matching approach (BM)*

Another traditional item selection approach is to select items with difficulty parameters (bs) as closely matching the estimated ability as possible. Though the simple minimization of the difference between the estimated ability and b is not identical to the maximization of item information, the BM approach has been used by many researchers because of its short computational time and its simple implementation algorithm. Moreover, the BM approach does not take into consideration of a parameter and thus would not place reliance on high a items (Chang & Ying, 1999).

Consequently, uneven item exposure distribution is not a problem in BM. Referring to Figure 2, for items with the same value of a, the item with $(\theta - b)$ equal to zero offers the largest information. Therefore, items from b-matching approach also provide a large amount of information on examinee's ability.

3.3. *Mixed approach*

In practice, CAT practitioners strive to balance the need for measurement accuracy, item security and item pool utilization. Leung *et al.* (2005) proposed a mixed item selection approach by capitalizing on the strengths of the traditional approaches, aimed at making good use of item bank without sacrifice of accuracy. The original mixed selection approach suggested to select items by BM for the first half (50%) of a test and then to switch to MI in the second half. The rationale for mixed approach is that at early stages, MI should not be used as the ability estimate is still far from the true value, or otherwise high a items would be selected and wasted. Instead, BM should be used at the early stages as it does not rely on high a items. High a items would be selected in the final stage by MI when ability estimate is close to the true value. In their study, the mixed approach was found to be as efficient as MI for CAT under a large set of non-statistical constraints.

This study addresses the need of developing effective continual and dynamic testing system for generating timely and reliable information on student achievement. In this study, the mixed approach is further examined. We attempt to investigate how this mixed approach would perform when there is no non-statistical constraint. Moreover, we vary the percentage of BM-selected items in order to explore any possible optimal trade-off between accuracy and item pool utilization. To facilitate the identification of a good design, we propose a new statistic for evaluating overall performance of each method on accuracy and item pool utilization.

4. Method

4.1. *Simulation study*

Three parameter logistic IRT model (3PLM) was used in this study. Under 3PLM, the probability for an examinee with ability parameter θ to correctly respond to an item with difficulty parameter b would follow Eq. (3.1). The Fisher information of this item for this examinee would follow Eq. (3.3). The MI approach selected the item with maximum Fisher information with respect to current ability estimate, while the b-matching approach selected the one with minimum absolute difference between item difficulty and current ability estimate. The value of the ability parameter was estimated using the Maximum Likelihood criterion (Lord, 1980). In this study, the abilities were assumed to follow normal distribution. So, maximum likelihood picks the value from the distribution that makes the data "more likely" than any other values of ability would make them. This picked value is then taken as the maximum likelihood estimate of examinee's ability.

Table 2. Descriptive statistics of different parameters used in the simulation study.

Parameter	Mean	s.d.	Max	Min
Ability (θ)	0.017	1.004	3.388	−3.135
Difficulty (b)	0.323	1.039	3.398	−3.444
Discrimination (a)	1.123	0.354	2.630	0.294
Pseudo guessing (c)	0.173	0.084	0.500	0.033

In this simulation study, tests were created by selecting items at the earlier stage using BM and selecting the remaining items using MI. Five different percentages of items in a test selected by MI were investigated: 0% (i.e. purely BM), 25%, 50%, 75% and 100% (purely MI). Two test lengths, 20-item and 40-item were simulated in order to see how test length interacts with performance. These ten combinations in the simulation design are by no means exhaustive. However, the results would provide a general picture on how the performance of the mixed approach varies with different proportion of MI-selected items across short test (20-item) and long test (40-item).

Under each of the 10 combinations of testing condition, 5,000 simulees with ability parameters randomly generated from a standard normal distribution, were tested using a retired pool of 700 calibrated mathematics items. The description statistics of these parameters are given in Table 2.

4.2. *Evaluation statistics*

Both general statistics and a newly proposed overall performance statistic were used to compare the performance of various CAT designs.

4.2.1. *General statistics*

Eight general statistics were used to reflect the performance of CAT algorithms in different aspects. These include:

(1) Average bias: $\frac{1}{n}\sum_{i=1}^{n}(\hat{\theta}_i - \theta_i)$, where θ_i and $\hat{\theta}_i$ denote the true and estimated ability of examinee i respectively, and n denotes the total number of examinees.
(2) Mean Squared Error (MSE): $\frac{1}{n}\sum_{i=1}^{n}(\hat{\theta}_i - \theta_i)^2$. A smaller MSE estimate would indicate a more efficient item selection method.
(3) Correlation: $\frac{n\sum_i \theta_i\hat{\theta}_i - \sum_i \theta_i \sum_i \hat{\theta}_i}{\sqrt{n\sum_i \theta_i^2 - (\sum_i \theta_i)^2}\sqrt{n\sum_i \hat{\theta}_i^2 - (\sum_i \hat{\theta}_i)^2}}$, the Pearson Correlation Coefficient between true and estimated abilities. The higher the correlation coefficient, the more reliable the test result is for decision making.
(4) Number of over-exposed items: number of items with exposure rate larger than 0.2, a commonly used cut-off value (see, e.g. Hau & Chang, 2001; Schaeffer *et al.*, 1995). Exposure rate of an item is the ratio of the number of times it is administered to the total number of tests. Too many over-exposed items

may cause security problem. Item selection methods with smaller numbers of over-exposed items are considered more desirable as regards to item security.

(5) Number of under-utilized items: number of items with exposure rate less than 0.02. If there are too many items with low exposure rates, then the item pool is not fully utilized. This phenomenon challenges the cost effectiveness of developing the items and hence the appropriateness of the item selection method.

(6) Scaled χ^2-statistics (χ^2): $\sum_{j=1}^{N} \frac{(er_j - L/N)^2}{L/N}$, where L and N represent test length and item pool size respectively, and er_j denotes the exposure rate of item j. It is an indicator for the efficiency of overall item pool usage. The smaller the χ^2, the more balanced is the pool utilization.

(7) Test overlap rate: the item-overlap rate was computed in the present study by (a) first counting the number of common items for each of the $n(n-1)/2$ pairs of examinees, (b) adding up all the counts in the total of $n(n-1)/2$ pairs, and (c) dividing the total count by $Ln(n-1)/2$. The higher the item overlap rate, the greater the damage to the test validity due to information sharing among examinees who take the test at different occasions (Leung *et al.*, 2005).

(8) Maximum item exposure rate: the overall maximum exposure rate of individual items across all tests. If an item has a high exposure rate, then it has an increased risk of being known by prospective examinees, which in turn threatens test security.

The first three statistics reflect performance in recovering the unknown ability parameters, while the rest indicate item security and item pool utilization.

4.2.2. *Overall performance*

The ideal value of MSE is 0, meaning that the ability estimate is exactly the same as the true value. Similarly, the ideal value of χ^2 is also 0, meaning that each item has the same exposure rate and the pool is fully utilized. Therefore, the ideal situation for accuracy and item pool utilization is that both MSE and χ^2 are zero. On a plot of MSE against χ^2, the farther away a point of the two statistics from the origin, the worse the overall performance of the item selection method is. Based on this argument, we propose an index for indicating overall performance which is measured by the scaled distance from (0, 0). Let the maximum observed value of χ^2 be x^* while that of MSE be $y*$. We propose a new statistic d, the weighted scaled distance of (x, y) from (0, 0) for comparing overall performance. And d can be expressed as

$$d = \sqrt{w_1(x/x^*)^2 + w_2(y/y^*)^2}, \tag{4.1}$$

where w_1 and w_2 are the weights that testing organizations can place on χ^2 (item pool utilization) and MSE (accuracy) respectively based on their needs. The smaller this value, the better is the overall performance of the corresponding CAT design.

It is true that other factors such as overlap rate, number of over-exposed items and number of under-utilized items are important for comparing performance of

item selection method. Nevertheless, better item pool utilization (smaller χ^2) means that more items are evenly administered to examinees. Consequently, the average number of overlap items, the number of under-utilized items and the number of over-exposed items would drop down. Therefore, putting the two key factors, MSE and χ^2, in the computation of weighted scaled distance provides an appropriate and parsimonious statistic for comparing overall performance.

Testing organizations may place their own weights on accuracy and item pool utilization. To one extreme, when accuracy is the utmost concern, w_2 can be put to 1 and w_1 be put to 0. In such case, Eq. (4.1) would become

$$d = y/y^*. \tag{4.2}$$

In Eq. (4.2), the best overall performance occurs when y (MSE) is minimized, as which in turn minimizes d (scaled distance).

To other extreme, some e-testing programs may like to use their caliabrated items as many as possible. For them, item pool utilization may be a major concern. They may set w_2 as 0 and w_1 as 1. In such case, Eq. (4.1) would become

$$d = x/x^*. \tag{4.3}$$

The best overall performance comes with the smallest scaled distrance d, which can be attained when x (χ^2 value) is minimized. This study imitates the scenario that testing programs set both weights to 1, putting both accuracy and item pool utilization criteria in equally important priorty.

5. Results

5.1. *Effect of ratio of items selected by MI on performance*

Table 3 lists the summary of the evaluation statistics of different CAT settings after tests were simulated. In all the settings, the estimated biases were all close to zero, showing that there were no systematic errors in ability estimation. As the correlation coefficient was significant and high (ranging from 0.913 to 0.983), the ability estimates from all item selection methods are reliable for making decisions.

In general, regardless of test length, statistics concerning item usage such as the number of over-exposed and under-utilized items, χ^2, test overlap rate and maximum exposure rate increase with the percentage of item selected by MI, showing that item pool utilization and item security are worsened.

In both 20-item and 40-item tests, the 100% MI item selection method hit the maximum exposure rate up to over 0.745 and 0.765 respectively. This will definitely cause security alarm as those highly over-exposed items were administered to 3 out of 4 examinees. The content of these over-exposed items would soon be known to the subsequent cohorts of examinees. Also, the overlap rates in this method are the highest (0.299 for 20-item and 0.309 for 40-item). It means that on average there were about 6 common items in any two 20-item tests and 12 common items in any two 40-item tests. The χ^2 values in this method are also the highest, indicating very

Table 3. Performance of the 10 CAT designs.

n^*	%MI	Bias	MSE	corr	nOv†	nUn†	χ^2	OvL‡	maE^	Dist§
20	100	−0.001	0.072	0.966	38	590	196.6	0.299	0.745	1.063
20	75	0.002	0.085	0.961	32	583	140.5	0.229	0.571	0.831
20	50	0.003	0.101	0.954	24	502	95.6	0.165	0.587	0.701
20	25	0.012	0.124	0.945	11	394	54.3	0.106	0.590	**0.678**
20	0	0.010	0.200	0.913	2	307	28.3	0.069	0.576	1.010
40	100	0.002	0.037	0.983	82	495	169.7	0.309	0.765	0.883
40	75	0.001	0.038	0.982	71	413	121.5	0.231	0.586	0.647
40	50	−0.001	0.042	0.980	48	249	82.0	0.174	0.579	0.468
40	25	−0.002	0.052	0.976	28	149	46.5	0.123	0.585	**0.350**
40	0	−0.001	0.069	0.968	2	84	19.1	0.084	0.573	0.358

Note: $^*n$: test length.
†nOv/ nUn: number of overexposed and underutilized items.
‡OvL: overlap rate.
^maE: maximum exposure rate.
§Dist: scaled distance.

poor item utilization. In fact, there were 590 (495) under-utilized items and 38 (82) over-exposed items after five-thousand 20-item (40-item) tests were finished.

When compared with the 100% MI method, the mixed item selection method employing 75% MI and 25% BM has better performance in item pool utilization and item security. It has lower χ^2 value, lower test overlap rate, and fewer numbers of under-utilized and over-exposed items. This trend continues when the percentage of MI selected items decreases.

In terms of accuracy, MSE decreases as the percentage of MI-selected items increases, indicating an improvement in accuracy. For both 20-item and 40-item tests, 100% MI CAT design yielded the smallest MSE, meaning that it was the best in accuracy. The 100% BM had the highest MSE while the results for mixed item selection methods varied between these two ends.

5.2. *Overall performance*

Figure 3 is the plot of MSE against scaled χ^2 statistics for the 10 settings in Table 3. The plot indicates that there was tradeoff between measurement accuracy and item pool utilization. As mentioned earlier, the closer a point to the origin (0, 0), the better is the overall performance. The tradeoff of MSE was slight when test became longer; in that case a great reduction in χ^2 only resulted in a small increase in MSE. This means that there would be a great improvement in item pool utilization with a slight sacrifice in measurement efficiency.

It was discovered that, regardless of test length, the setting "first 75% of the items selected by BM" was the setting resulted in minimal scaled distance from the origin. For 75% BM mixed item selection method, the scaled distance reached the minimum of 0.350 for 20-item test and 0.678 for 40-item test. In this method,

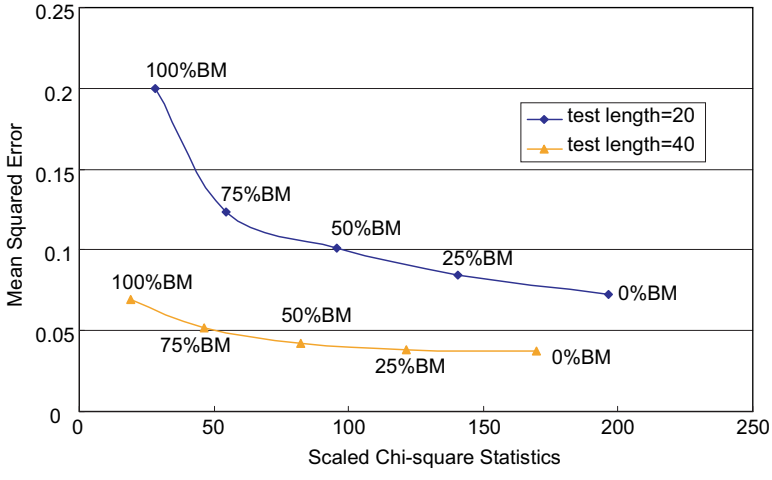

Figure 3. A plot of Mean Squared Error against scaled χ^2-statistics for ten different settings.

the correlation coefficients for the true abilities and their estimates were 0.945 (20-item test) and 0.976 (40-item test). The high correlation indicates that this method is able to provide reliable estimates of the abilities.

6. Conclusions and Discussion

With the advancement of computer technology and increasing availability of high-speed computers, computerized adaptive testing will take an important role in enhancing student learning. One attrative advantage of computerized adaptive testing that the conventional paper-and-pencil tests do not have is its potential of incorporating alternate item formats that may involve interactive graphics, sounds and video capture for realistic situations. These new item formats and testing environments can stimulate student thinking and arouse student motivation in learning.

Moreover, it is generally agreed that a standard set of items for all examinees is against a very fundamental principle of education: catering for individual needs. As computerized adaptive testing administers items to individuals according to their abilities, it can deliver tailor-made tests. The items for each individual are neither too difficult nor too simple for any particular examinee. As such, computerized adaptive testing can support educational systems to meet individual needs.

To date, teachers spend a lot of time on test preparation, marking, giving responses, and reporting. Therefore, it is a very meaningful challenge and important task to develop a more efficient and dynamic testing system for enhancing learning by providing timely quality information on student achievement. When each student can conveniently evaluate individual progress, he or she would have higher motivation to regulate his or her learning pace.

With the increasing availability of powerful computers and the advancement of statistical theories, tailoring tests to individual's ability has become more easily realized. The major issue is on developing a good item selection method which can make good use of item pool and provide reliable estimation.

In the mixed item selection approach, the increase in percentage of items selected by MI results in higher measurement accuracy but poorer item pool utilization. This pattern of trade-off is a consequence of the item selection strategies of MI and BM. As a test proceeds, the fluctuation of the value of intermediate ability estimate decreases and the estimated value converges to the true value. An item provides larger Fisher information if its difficulty is close to the ability and its discrimination is high. Therefore MI tends to select from a small number of high a items, resulting in a skewed item usage pattern. In BM, $(b - \theta)$ constitutes the sole criterion in selecting and thus is less selective compared to MI. However, less informative items may be selected by BM.

The trading off for measurement accuracy by sacrificing item pool utilization is less acute for long test, as accuracy is too high to be further increased by increasing information.

A plot of MSE against scaled χ^2 statistics can provide us knowledge about the overall performance of a CAT algorithm. Under the same test length, the algorithm with corresponding dot on the plot closest to the origin (0, 0) would have the best balance between accuracy and item pool utilization. In the current study, the best overall performance of mixed item selection approach was obtained when the ratio of the number of items selected by BM to that by MI is close to 3:1.

In this study, different designs of mixed item selection approaches are compared in terms of measurement accuracy and item utilization efficiency. Mixed item selection approach has demonstrated its strengths in allowing testing organizations more flexibility in optimizing the competitive goals of accuracy and item pool utilization, simply by changing the ratio of items selected with each item selection approach (maximum information/b-matching). It can result in satisfactory measurement accuracy and efficiency in item utilization according to the needs of individual testing programs.

The results of the current study by no means represent the relative performance of different types of mixed selection methods under all situations. The performance of a mixed item selection may depend on the distribution of item and ability parameters. For example, if the average discrimination level of the item pool is low, a higher proportion of items selected by MI may be preferred. Nevertheless, mixed item selection approach is shown to have great potential to introduce flexibility into many existing CAT algorithms. CAT practitioners are advised to do prior simulation studies to obtain an optimal mixed item selection method with their own set of items and anticipated distribution of examinee abilities, before deciding which one should be used in the real e-testing.

This study is significant in at least two aspects. First, the empirical data of the study have shown that there always exists some good mixed item selection

methods in between the pure *b*-matching method and the pure maximum information method. Second, this study formulates a new statistic to help practitioners identify a good selection method according to their needs. Future research may pursue on this line to further investigate other possible combined statistics to evaluate the overall performance of e-testing designs.

Acknowledgments

This paper is an extended version of the paper presented at The 17th International Conference on Computers in Education, ICCE 2009. The author thanks the anonymous reviewers for their helpful comments and both The Hong Kong Institute of Education and the General Research Fund for their financial support to the study. (Project code: HKIEd841909)

References

Armstrong, R., Belov, D., & Weissman, A. (2005). Developing and assembling the law school admission test. *Interfaces*, *35*, 140–151.

Beevers, C. E., McGuire, G. R., Stirling, G., & Wild, D. G. (1995). Mathematical ability assessed by computer. *Computers & Education*, *25*(3), 123–132.

Boekaerts, M., & Corno, L. (2005). Self-regulation in the classroom: A perspective on assessment and intervention. *Applied Psychology: An International Review*, *54*(2), 199–231.

Chang H. H., & Ying, Z. L. (1999). A-stratified multistage computerized adaptive testing. *Applied Psychological Measurement*, *23*(3), 211–222.

Conejo, R., Guzmán, E., Millán, E., Trella, M., Pérez-De-La-Cruz, J., & Rios, A. (2004). SIETTE: A web-based tool for adaptive testing. *International Journal of Artificial Intelligence in Education*, *14*(1), 29–61.

Hambleton, R., & Swaminathan, H. (1985). *Item Response Theory: Principles and Applications*. Boston, MA: Kluwer Nijhoff.

Hau, K. T., & Chang, H. H. (2001). Item selection in computerized adaptive testing: Should more discriminating items be used first? *Journal of Educational Measurement*, *38*(3), 249–266.

Leung, C. K. (1998). Computerized adaptive testing as a means for mathematics assessment. *EduMath*, *7*, 21–27.

Leung, C. K., Chang, H. H., & Hau, K. T. (2005). Computerized adaptive testing: A mixture item selection approach for constrained situations. *British Journal of Mathematical & Statistical Psychology*, *58*(2), 239–257.

Lord, M. F. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale NJ: Lawrence Erlbaum Associates.

Schaeffer, G., Steffen, M., Golub-Smith, M., Mills, C. N., & Durso, R. (1995). *The Introduction and Comparability of the Computer Adaptive GRE General Test* (Research Report 95–20). 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*(1), 57–75.

Straetmans, G. J. J. M., & Eggen, T. J. H. M. (1998). Computerized adaptive testing: What it is and how it works. *Educational Technology*, *38*(1), 45–52.

Ueno, M. (1998). The open testing system. In *The Proceedings of Open Learning 98* (pp. 299–307).

van der Linden, W. J. (1998). Bayesian item selection criteria for adaptive testing. *Psychometrika*, *63*(2), 201–216.

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., Mislevy, R. J., Steinberg, L., & Thissen, D. (1990). *Computerized Adaptive Testing: A Primer*. Hillsdale NJ: Lawrence Erlbaum Associates.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement*, *6*(4), 473–492.

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, *21*(4), 361–375.