Applied Psychological Measurement
2018, Vol. 42(6) 499–515
© The Author(s) 2018
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0146621617752990
journals.sagepub.com/home/apm

What Information Works Best?: A Comparison of Routing Methods

Halil Ibrahim Sari[1] and Anthony Raborn[2]

Abstract

There are many item selection methods proposed for computerized adaptive testing (CAT) applications. However, not all of them have been used in computerized multistage testing (ca-MST). This study uses some item selection methods as a routing method in ca-MST framework. These are maximum Fisher information (MFI), maximum likelihood weighted information (MLWI), maximum posterior weighted information (MPWI), Kullback–Leibler (KL), and posterior Kullback–Leibler (KLP). The main purpose of this study is to examine the performance of these methods when they are used as a routing method in ca-MST applications. These five information methods under four ca-MST panel designs and two test lengths (30 items and 60 items) were tested using the parameters of a real item bank. Results were evaluated with overall findings (mean bias, root mean square error, correlation between true and estimated thetas, and module exposure rates) and conditional findings (conditional absolute bias, standard error of measurement, and root mean square error). It was found that test length affected the outcomes much more than other study conditions. Under 30-item conditions, 1-3 designs outperformed other panel designs. Under 60-item conditions, 1-3-3 designs were better than other panel designs. Each routing method performed well under particular conditions; there was no clear best method in the studied conditions. The recommendations for routing methods in any particular condition were provided for researchers and practitioners as well as the limitations of these results.

Keywords

computerized multistage testing, information, routing method

Many test administration models are used to measure student success in the educational and psychological measurement. The commonly used method is the linear paper-and-pencil test. Even though there are some practical advantages of this kind of test, such as low cost of test administration and flexible scheduling, it is criticized for multiple reasons, the most pressing being for producing low measurement precision in student scores (Yan, von Davier, & Lewis,

[1] Kilis 7 Aralik University, Turkey
[2] University of Florida, Gainesville, USA

Corresponding Author:
Halil Ibrahim Sari, Measurement and Evaluation in Education Program, Muallim Rifat Education Faculty, Kilis 7 Aralik University, Kilis 79100, Turkey.
Email: hisari@kilis.edu.tr

2014). Adaptive testing has been proposed to overcome some of the problems associated with paper-and-pencil tests.

The most widely known adaptive test is computerized adaptive testing (CAT). In this testing approach, the exam is administered on computer, and depending on the current performance on an item, the next item is tailored to test taker until the exam ends (van der Linden & Glas, 2000). This means that there is an item-level adaptation in CAT which is achieved by some item selection algorithm. Even though CAT produces the best measurement accuracy among all available test administration models (Yan et al., 2014), it is also criticized in terms of computational complexity, high cost of test administration, item review, and item skipping. Due to item-level adaptation, each time the computer algorithm calculates one's latent score (i.e., theta estimate) and selects the most informative item from the item bank based on this estimate, and after he or she responds to that item, it reestimates the theta estimate and selects another item. If the test length is 60, there must be 59 adaptation points for an examinee during the test (e.g., first item which is typically at medium difficulty level is given from the item pool so, not adapted). Furthermore, in a CAT administration, examinees cannot revisit the previous items during test nor skip any items.

Computerized multistage testing (ca-MST) is another adaptive test administration model and one of the proposed alternatives to the CAT. As its name implies, ca-MST is comprised of multiple stages within which are item groups at different difficulty levels called modules. There is typically one module in Stage 1 which is called the routing module. After each examinee receives this routing module in Stage 1, a computer algorithm tailors one of the available modules in Stage 2 depending on the performance on this module (Luecht & Sireci, 2011). This basically means that there is a module-level adaptation, leading to fewer adaptation points in ca-MST than in traditional CAT. In terms of item review, item skip, computational intensity, measurement precision, and test security, the ca-MST lies in between linear pencil and paper tests and CAT.

There are some important elements in ca-MST. One of them is the routing method—that is, the algorithm used to decide the next module an examinee should receive. Routing method is analogous to the item selection method in CAT and determines the efficiency of a ca-MST, thereby affecting the measurement accuracy of theta estimates at the end of the test (Lord, 1980). If the routing method is not carefully chosen, misrouting would likely occur and one might draw an undesired or unexpected pathway during the test which harms the adaptability feature of ca-MST by limiting the possible theta estimates and increasing the standard error of measurement.

There are two popular sets of routing methods used in ca-MST framework. One is the number-correct (NC) method, which is based on an NC score. This method routes examinees to the next module based on the number of correct responses they provided in the current module. That is, for example, if an examinee gets five items or less out of 15, he or she receives an easy (lower item difficulty) module; between six items and 10 items, receives a medium module; 11 items or more, receives a hard module in the next stage. One can refer to Zenisky, Hambleton, and Luecht (2010) for more details in determining the cutoff points in this method. The NC is a fairly straightforward method and produces comparable or slightly better results in theta estimates than information-based routing method (Weissman, Belov, & Armstrong, 2007). However, NC works best under the Rasch and one-parameter logistic (1PL) models as NC is a sufficient statistic for theta; for more complicated models, using NC can result in a loss of accuracy in the estimate as the items are not statistically equivalent (Lord, 1980). Even if the items in a module are at the similar difficulty level, potentially different discriminating power for the items means they do not necessarily provide equal information for the theta estimate. When estimating the final theta, the contribution of more discriminating items will be higher and so it matters which items one gets correct.

The other popular set is the information-based routing methods which are the main interest in this study. As its name implies, item response theory (IRT) information is an integral part of these methods. These types of routing methods consider test (module) level information functions when selecting the subsequent modules. That is, the computer estimates an examinee's theta level based on the module(s) he or she received and then selects the next module that maximizes information at her current theta estimate (Weissman et al., 2007). The most widely known information-based routing method is maximum Fisher information (MFI). This method has been widely used as an item selection method in CAT framework, and then adopted to ca-MST framework. MFI in the CAT framework finds the most informative item for provisional theta after each theta estimate and selects that individual item for the examinee. MFI in the ca-MST framework works similarly but the provisional theta is estimated after an examinee receives a module, and then the most informative module in the next stage is selected by examining cumulative item information across the available modules in the next stage (e.g., module-level information). Some studies have shown MFI to work more efficiently within the CAT framework as opposed to the ca-MST framework; however, it was still comparably efficient in ca-MST.

There are other item selection methods proposed for CAT applications. It is believed that these can also be used as a routing method in ca-MST applications by using a similar idea (e.g., examining module-level information). In this study, besides MFI, four additional item selection methods were selected and used them as a routing method in ca-MST framework. These are maximum likelihood weighted information (MLWI; Veerkamp & Berger, 1997), maximum posterior weighted information (MPWI; van der Linden, 1998), the Kullback–Leibler (KL; Chang & Ying, 1996), and posterior Kullback–Leibler (KLP) criterion (Chang & Ying, 1996). The main purpose of this study is to examine the performance of these methods when they are used as a routing method in ca-MST applications. It has already been shown that test length and panel structure of ca-MST play important roles in applying ca-MST, where panel structure refers to the number of stages and the number of modules within each stage (Luecht, 2000). Some researchers argue that having more stages and more modules in each stage allows more branching and is better (Luecht & Nungester, 1998), while others believe that having more than three stages and more than three or four modules in each stage does not meaningfully increase test outcomes (Patsula & Hambleton, 1999). To investigate these viewpoints, four different ca-MST panel structures and two test lengths were manipulated to explore the efficiency of these methods in ca-MST framework.

It should be emphasized that the choice of routing method depends on the purpose and consequences of the test (Zenisky et al., 2010). These five methods are chosen because they fulfill similar requirements for the purpose and consequence of a test and are therefore comparable under similar circumstances. In addition, information-based methods have been less commonly studied methods; this study aims to examine how these methods perform under various conditions to inform practitioners and researchers on each of their efficiencies in those conditions.

Background on Adaptive Selection Methods

Maximum information method (Lord, 1980) is the oldest and most widely used selection method (Veerkamp & Berger, 1997). In CAT framework, after one's provisional theta estimate is calculated based on the previously administered items, this method selects the next item that maximizes information at his or her current ability estimate (Weissman et al., 2007). However, MFI is calculated at a fixed theta value (see Equation 2). When it is used as item selection method, in early stages of CAT there is usually higher discrepancy between provisional and true thetas. Therefore, due to a combination of estimation error and higher initial standard errors of theta estimates, the most informative item is not always chosen. Veerkamp and Berger (1997) proposed MLWI in CAT which is the product of information function at fixed theta value and likelihood function. This method attempts to find most informative item over a range of theta values instead of a fixed value by weighting MFI using the theta estimation interval; therefore, some call it Fisher interval information (see Wang & Chen, 2004). On the contrary, van der Linden (1998) suggested MPWI, which is computed by multiplying the information function by a posterior theta distribution based on previously administrated items instead of likelihood function. These two methods are based on Fisher information, but after calculating the Fisher information, it is weighted by either the likelihood or posterior distribution for MLWI and MPWI, respectively. When used as routing methods in the ca-MST framework, the idea is exactly the same but the information will be calculated at the module level for all available modules at that stage, and then the one that maximizes module-level information is selected. The formulas for these three methods were briefly given in the appendix, and comprehensive explanations and more technical details of them can be found in S.-Y. Chen and Ankenmann (2004) and van der Linden and Glas (2000).

The item selection version of KL method was first introduced by Chang and Ying (1996) and is based on a log-likelihood ratio test. In the CAT framework, this method calculates the nonsymmetric distance between two likelihoods at two estimated trait levels, called KL information gain. KL is the ratio of two likelihood functions instead of a fixed value as in the MFI. Similar to MPWI, Chang and Ying proposed weighting KL information by multiplying the posterior density of the current ability estimate by KL information. This is called the KL with posterior information method (KLP). Similar to the additive property of MFI, it is also possible to calculate KL or KLP information gains at module level by summing item-level KL information gains. The idea in ca-MST framework is that the KL or KLP information for all available modules is first estimated at that stage and then the next module that maximizes the KL or KLP information is selected. One can refer to Chang and Ying (1996) and S.-Y. Chen and Ankenmann (2004) for more technical details. The formulas for these two item selection methods were presented in the appendix.

Method

Design Overview

In this study, five routing methods (MFI, MLWI, MPWI, KL, and KLP) were compared across several conditions including test length with two levels and ca-MST panel design with four levels. The two levels of test lengths were 30 items and 60 items. As shown in Figure 1, the four levels of ca-MST panel designs were 1-3, 1-2-2, 1-2-3, and 1-3-3, which have been commonly chosen in the ca-MST literature (see Schnipke & Reese, 1999; Zenisky, 2004). All manipulated conditions were fully crossed with one another. This resulted in 40 conditions (5 Routing Methods × 2 Test Lengths × 4 Panel Designs). For each condition, 100 replications were performed. For each replication, a set of 3,000 examinees were generated from a standard normal distribution. Within a single replication, the set of theta values were used in each of the 40 conditions. Both fixed and varied study conditions were detailed in following sections. The whole simulation process was completed in R Version 2.1.1 (R Development Core Team, 2009-2015).

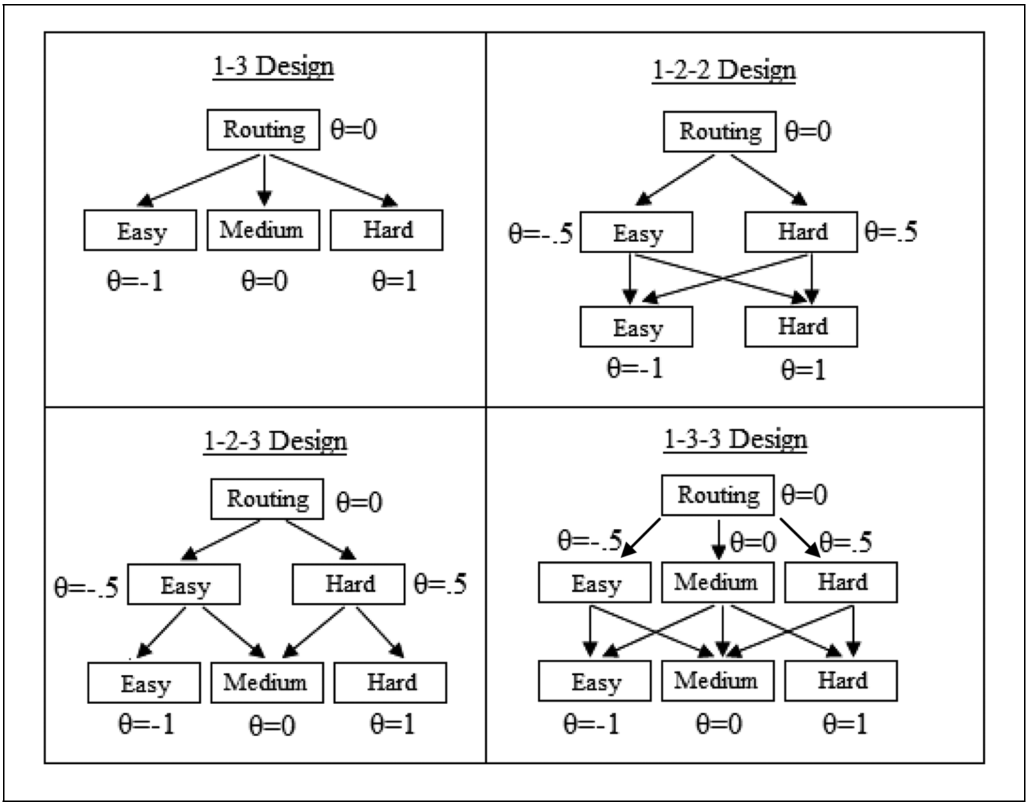


Figure 1. Studied ca-MST panel designs.
Note. ca-MST = computerized multistage testing.

Fixed Study Conditions

In this study, the item parameters were based on a real Armed Services Vocational Aptitude Battery (ASVAB) military test used in Armstrong, Jones, Li, and Wu (1996). As in the original item bank, the simulated item bank had 450 multiple choice items from four different content areas. The item parameters and number of items for each content area were given in Table 1. The target test length in the ASVAB was 30, and the distributions across the content areas were 10, 11, 4, and 5 for Content 1, Content 2, Content 3, and Content 4, respectively. Under 30-item test length condition, the same target numbers for each content area were used, while under 60-item condition all corresponding numbers were doubled. It should be noted that within each panel structure, the number of items and content distributions for the modules at the same stage were the same.

As explained before, a total of 3,000 examinees were generated from a normal distribution. The three-parameter logistic (3PL; Birnbaum, 1968) IRT model was used to generate item responses. The 3PL model defines the conditional probability of a correct response on item *i* for person *p* (X_{ip} = 1) as

$$P\left(X_{ip}=1\middle|\theta_p\right)=c_i+(1-c_i)\frac{e^{a_i(\theta_p-b_i)}}{1+e^{a_i(\theta_p-b_i)}}, \quad (1)$$

Table 1. Item Parameters of Each Content Area in the Item Bank.

Content area (number of items)	*a*		*b*		*c*	
	M	*SD*	*M*	*SD*	*M*	*SD*
Content 1 (*n* = 150)	1.079	0.409	−0.467	1.179	0.210	0.095
Content 2 (*n* = 165)	1.128	0.438	−0.154	1.033	0.200	0.104
Content 3 (*n* = 60)	1.092	0.538	−0.025	0.815	0.203	0.084
Content 4 (*n* = 75)	1.237	0.383	−0.014	0.678	0.162	0.080

where b_i is the difficulty parameter, a_i is the item discrimination, c_i is the pseudo-guessing parameter for item i, and θ_p is the latent trait for person p. The expected a posteriori (EAP; Bock & Mislevy, 1982) with a prior distribution of $N(0, 1)$ was used for both provisional and final ability estimates across all simulation conditions.

Varied Conditions

The five information-based routing methods were MFI, MLWI, MPWI, KL, and KLP distribution. These methods are very popular and widely studied item selection methods in CAT (Barrada, Olea, Ponsoda, & Abad, 2010; S.-Y. Chen & Ankenmann, 2004; S.-Y. Chen, Ankenmann, & Chang, 2000; Veerkamp & Berger, 1997). The purpose of manipulating this condition was to examine the performance of routing method across the conditions.

The four panel structures were 1-3, 1-2-2, 1-2-3, and 1-3-3. These panel designs were given in Figure 1. The 1-3 panel design has two stages, but the others have three stages. The main difference between 1-2-2 design and 1-2-3 design is that 1-2-3 design has a third module in Stage 3. The main difference between 1-2-3 design and 1-3-3 design is that 1-3-3 design has a third module in Stage 2. The purpose of manipulating multiple panel designs was to explore whether performance of routing method is affected by the panel complexity. As illustrated in this figure, extreme jumps were not allowed to prevent aberrant response patterns as done in many studies (Luecht, Brumfield, & Breithaupt, 2006). This means that individuals could only move to a module in the next stage that has a difference of up to one level of difficulty from the module in the current stage and only affected the paths for the 1-2-3 and 1-3-3 designs. Finally, the test lengths were 30-item and 60-item test lengths, which would represent short and long test lengths, respectively. Again, all modules at any stage had the same number of items within each panel design. The content distributions within the modules in each panel structure were the same under both test lengths. The number of items in each module for 60-item test length was twice that of the 30-item test.

Test Assembly

In all ca-MST designs, three nonoverlapping essentially parallel panels were generated from the item bank. The total test information across the panel designs and test length conditions were given in Supplementary Figure 1. Pathway-level (e.g., routing-easy-easy or routing-hard-easy) information function was also provided for all possible pathways across the panel designs, test lengths (see Supplementary Figures 2, 3, 4, and 5). Multiple panels in ca-MST were created, aiming to hold the maximum panel exposure rates at 0.33. After the panels were built, 3,000 examinees were randomly assigned to the panels with 1,000 examinees in each panel. The IBM CPLEX program (ILOG, 2006) was used to build modules. First items were clustered into different modules, then modules (e.g., group of items) were randomly and simultaneously assigned to the three panels. The automated test assembly finds a solution to maximize the IRT information function at a fixed theta point; denote θ_0 as the fixed theta point. The binary decision variable is first defined, x_i (e.g., $x_i = 0$ means item i is not selected from the item bank, $x_i = 1$ means item i is selected from the item bank). The information function that is to be maximized is

$$I(\theta_0) = \sum_{i=1}^{N} I(\theta_0, \xi_i)x_i, \tag{2}$$

where ξ_i represents the item parameters of item i (e.g., a, b, c parameters). There were four content areas (e.g., C_1, C_2, C_3, and C_4). For example, in 1-3 and 30-item condition, for the easy module in a panel, five, six, two, and two items were needed from the four contents, respectively. This means that 15, 18, six, and six for a total of 45 easy module items for the three panels were needed. These 45 easy items were pulled at a time and then randomly and simultaneously assigned to the three panels so that each easy module in each panel has 15 items. The automated test assembly is modeled to maximize

$$\sum_{i=1}^{N} I(\theta_0, \xi_i)x_i, \tag{3}$$

subject to

$$\sum_{i \in C_1} x_i = 15, \tag{4}$$

$$\sum_{i \in C_2} x_i = 18, \tag{5}$$

$$\sum_{i \in C_3} x_i = 6, \tag{6}$$

$$\sum_{i \in C_4} x_i = 6, \tag{7}$$

$$\sum_{i=1}^{N} x_i = 45, \tag{8}$$

and

$$x_i \in (0, 1), \; i = 1, \ldots, N, \tag{9}$$

which puts constraints on C_1, C_2, C_3, and C_4; the total test length; and the range of decision variables, respectively. The N is the total number of items in the item bank. The content distributions in the modules across the design conditions were given in Supplementary Table 1. The numbers under 60-item conditions were doubled. The test assembly that models the other conditions was modeled similarly.

In all panel conditions, the fixed theta scores were given in Figure 1. As can be inferred from Figure 1, the items in the routing modules and medium modules were chosen from medium difficulty items (i.e., items that maximize information function at theta point of 0). In the three-stage panel designs, the fixed theta scores for the second stages were $\theta = -0.5$ and $\theta = 0.5$ for easy and hard modules, respectively. In all panel designs, the fixed theta scores for the last stages were $\theta = -1$ and $\theta = 1$ for easy and hard modules, respectively. The purpose of choosing these theta scores for module difficulty was to have more information at the end of the test as done in P.-H. Chen, Chang, and Wu (2012).

Evaluation Criteria

The results of the simulation were evaluated with two sets of statistics: (a) overall results and (b) conditional results as evaluated in similar studies (see Zenisky, 2004). For overall statistics, mean bias, root mean square error (RMSE), correlation between estimated and true theta ($\rho_{\hat{\theta}\theta}$), and module exposure rates as percentages (e.g., module utilization) were computed from the simulation results. Overall statistics are calculated within each replication for the 3,000 examinees, then averaged across all replications. Mean bias was calculated as

$$\bar{e} = \frac{\sum_{j=1}^{N} \left(\hat{\theta}_j - \theta_j\right)}{N}. \tag{10}$$

RMSE was calculated as

$$\text{RMSE} = \sqrt{\frac{\sum_{j=1}^{N} \left(\hat{\theta}_j - \theta_j\right)^2}{N}}. \tag{11}$$

Correlation between true and estimated theta values was calculated as

$$\rho_{\hat{\theta}_i, \theta_j} = \frac{\text{cov}\left(\hat{\theta}_j, \theta_j\right)}{\sigma_{\hat{\theta}_j} \sigma_{\theta_j}}. \tag{12}$$

The conditional statistics represent the calculations at a particular range of estimated theta value. For conditional results, conditional standard error of measurement (CSEM), conditional absolute bias, and conditional RMSE were calculated between $\theta = -2$ and $\theta = 2$, with the width of the θ interval at 0.1, resulting in 41 fixed theta ranges. The number of examinees within each range of theta values ranged from approximately 10 (values −2.0 to −1.8 and 1.8 to 2.0) to 50 (values −2.0 to −1.8 and 1.8 to 2.0) within a single replication.

Results

Overall Results

Under the 30-item conditions, the mean bias, RMSE, and correlations for all routing methods and ca-MST panel designs were approximately .012, .33, and .94, respectively. Under 60-item condition, the mean bias, RMSE, and correlations for all routing methods and ca-MST panel designs were approximately .00, .27, and .96, respectively. The main finding was that type of routing method and ca-MST panel complexity did not affect the overall outcomes. However, increasing test length from 30 items to 60 items improved the outcomes.

As discussed before, three panels were generated in all ca-MST design conditions, and 3,000 examinees have been randomly assigned to one of the three panels. This means that the

Table 2. Conditional Exposure Rate Percentages Across the Conditions.

		30 items						60 items					
		Stage 2			Stage 3			Stage 2			Stage 3		
Design	Routing	E	M	H	E	M	H	E	M	H	E	M	H
1-3	MFI	37	34	28	—	—	—	35	34	31	—	—	—
	KL	46	21	33	—	—	—	41	25	34	—	—	—
	KLP	40	38	22	—	—	—	37	32	31	—	—	—
	MLWI	41	27	32	—	—	—	38	28	34	—	—	—
	MPWI	38	35	27	—	—	—	35	34	31	—	—	—
1-2-2	MFI	55	—	45	55	—	45	52	—	48	53	—	47
	KL	55	—	45	55	—	45	53	—	47	53	—	47
	KLP	55	—	45	55	—	45	53	—	47	53	—	47
	MLWI	55	—	45	55	—	45	53	—	47	53	—	47
	MPWI	55	—	45	55	—	45	53	—	47	53	—	47
1-2-3	MFI	55	—	45	35	35	30	54	—	46	37	29	34
	KL	57	—	43	39	29	31	55	—	45	39	26	35
	KLP	57	—	43	36	33	29	55	—	46	38	28	34
	MLWI	56	—	44	38	30	32	55	—	45	40	25	35
	MPWI	57	—	43	36	34	30	55	—	45	38	28	34
1-3-3	MFI	51	9	40	39	30	31	55	0	45	45	18	37
	KL	55	4	41	42	24	34	55	1	44	46	16	38
	KLP	53	7	40	40	28	32	56	0	44	46	16	38
	MLWI	53	6	41	41	25	34	55	0	45	47	14	39
	MPWI	53	7	40	40	28	32	56	0	44	47	15	38

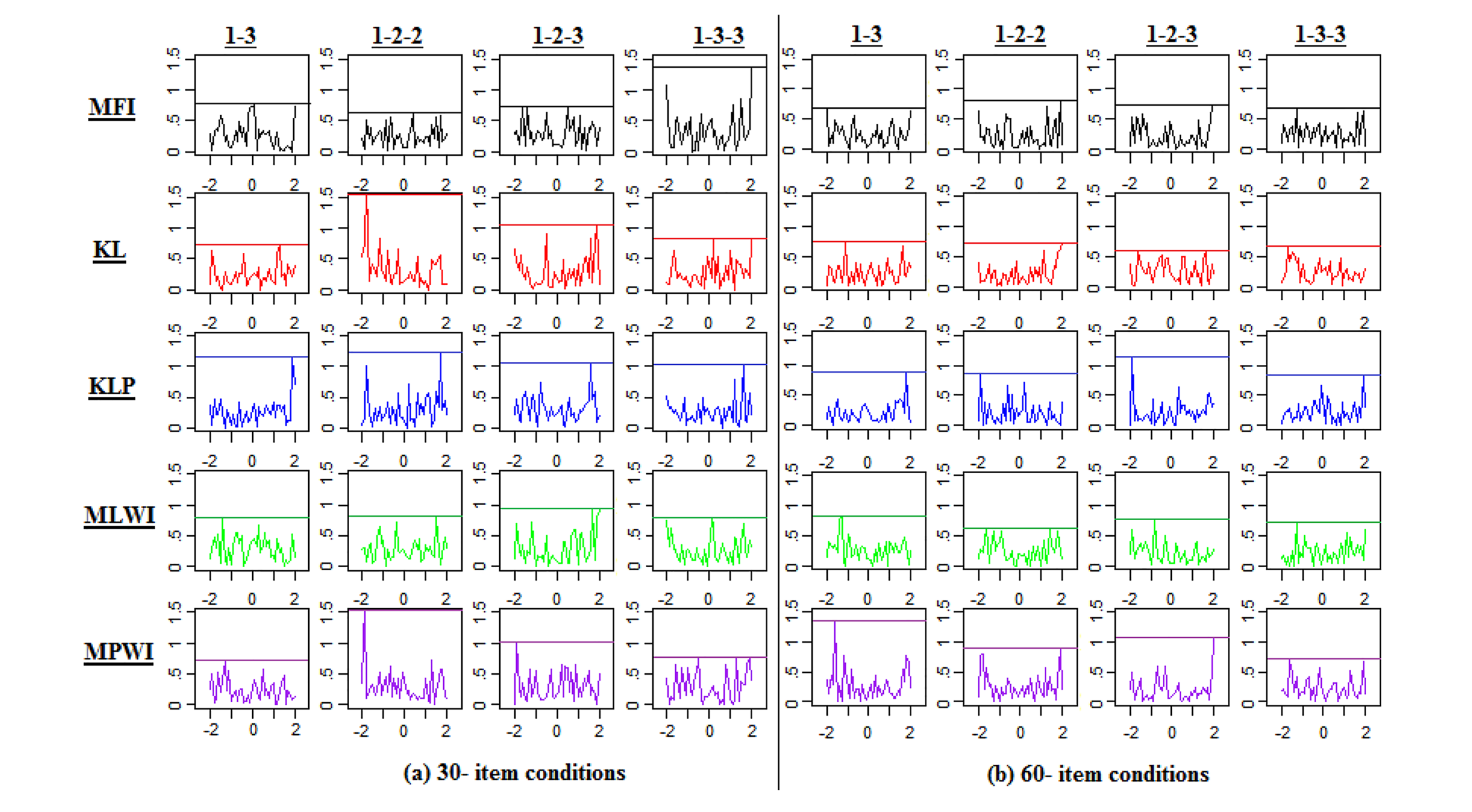
Note. E = easy; M = medium; H = hard; MFI = maximum Fisher information; KL = Kullback–Leibler; KLP = posterior Kullback–Leibler; MLWI = maximum likelihood weighted information; MPWI = maximum posterior weighted information.

exposure rate for a panel was 1/3. As the routing modules in Stage 1 were seen by all examinees that were assigned to that panel, exposure rate for routing modules was 100%. However, depending on the ca-MST design (e.g., two- or three-stage design), the modules in Stage 2 or Stage 3 were seen by fewer number of examinees. The conditional module exposure rates across all studied conditions, expressed as percent exposure within a stage, were presented in Table 2.

The first finding was that regardless of test length, ca-MST design and routing method, exposure rates for easy modules were higher than other modules. Due to lack of medium module, this was more obvious at Stages 2 and 3 in 1-2-2 design and at Stage 2 in 1-2-3 design. The second finding was that regardless of test length all routing methods yielded very similar exposure rates, especially under 1-2-2 design. The third finding was that the exposure rates for medium modules at Stage 2 were visibly lower than other modules in 1-3-3 design for all routing modules, and this was even lower when the test length 60. Within the same design, the exposure rates for Stage 3 medium modules were much higher for all routing methods across the both test lengths. Overall, in terms of conditional exposure rates, this type of design affected the findings much more than the other study conditions.

Conditional Results

Conditional absolute bias. The results of conditional absolute bias across the five routing methods and four panel designs and two test lengths were given in Figure 2. Based on this figure, the



(a) 30- item conditions

(b) 60- item conditions

Figure 2. Conditional absolute biases across all study conditions.

Note. MFI = maximum Fisher information; KL = Kullback–Leibler; KLP = posterior Kullback–Leibler; MLWI = maximum likelihood weighted information; MPWI = maximum posterior weighted information.

first finding was that the same routing method produced more varied biases under the three-stage designs than the two-stage design. This was more obvious especially for extreme theta values. The second finding was that as the number of items increased the absolute bias tended to be similar or lower with the exception of MPWI, which had higher bias within the 1-3 and 1-2-3 panel designs and similar variability overall. The final finding was that the 1-3 panel design was more stable (i.e., less variable) under the 30-item condition while the 1-3-3 design was more stable under the 60-item condition across all routing methods. Based on this figure, the MFI and MLWI methods had the most stable conditional and absolute bias.

Conditional RMSE. The results of conditional RMSEs across the five routing methods and four panel designs and two test lengths were given in Figure 3. These results are very similar to those found in Figure 2 in that the RMSE is more stable with 60-items as compared with 30 items and the MPWI routing method tended to have more variability and higher RMSE within the 60-item 1-3 as compared with the 30-item condition. Again, this was more obvious especially for more extreme theta values. In general, the variability of RMSE is similar to the variability of the absolute bias with the exception that, if the MPWI method is ignored, the 1-3 panel design had lower variability with more items. Even in the MPWI 1-3 conditions, looking at the middle theta values show that the RMSE is similar or lower with 60 items. Based on these results, the MFI and MLWI methods had the most stable RMSE under both item conditions; additionally, the KL had stable RMSE values under the 60-item condition.

CSEM. The results of CSEMs across the five routing methods and four panel designs and two test lengths were given in Figure 4. The main finding was that test length played much more impact than other study conditions on the conditional standard errors. As seen in Figure 3, as the test length increased standard error of measurements across the estimated theta scores decreased. The second finding was that standard error of measurements were always lower around theta = 0 and higher at more extreme theta points, which is an expected result given that the routing module in each condition maximized information at theta = 0. This was more obvious under 30-item conditions in 1-2-2 and 1-2-3 ca-MST panel designs.

Discussion and Limitations

The routing method is an integral part of a ca-MST application because it achieves adaptability feature of ca-MST and determinates efficiency of it (Luecht et al., 2006). So far, a noninformation routing method of NC has been the most commonly preferred routing method in the literature. It is also currently used routing rule in the revised Graduate Record Examinations (r-GRE). Admittedly, the NC routing rule is fairly straightforward method and understandable to the test takers as well. The focus in this study was information-based routing methods in ca-MST applications. In this study, some popular item selection methods proposed in CAT framework were selected and used them as routing methods in the ca-MST framework. Although there have not been many information-based routing methods used in ca-MST and the most widely used is MFI, the MFI, KL, KLP, MLWI, and MPWI methods were included and their efficiency was examined under varying conditions including different panel designs and test lengths. This work aims to emphasize and empirically support the best choice of routing method in a particular condition. In the literature, using these CAT procedures in the ca-MST framework was not discussed; it is believed that this study shows their relative efficiency and provides evidence that they can also be used as routing methods in testlet or module-level applications (e.g., ca-MST).

Based on the findings, on the overall results all routing methods were equally efficient in all panel designs within each test length condition. This means that type of routing method and panel design did not affect the overall findings; however, test length affected the overall

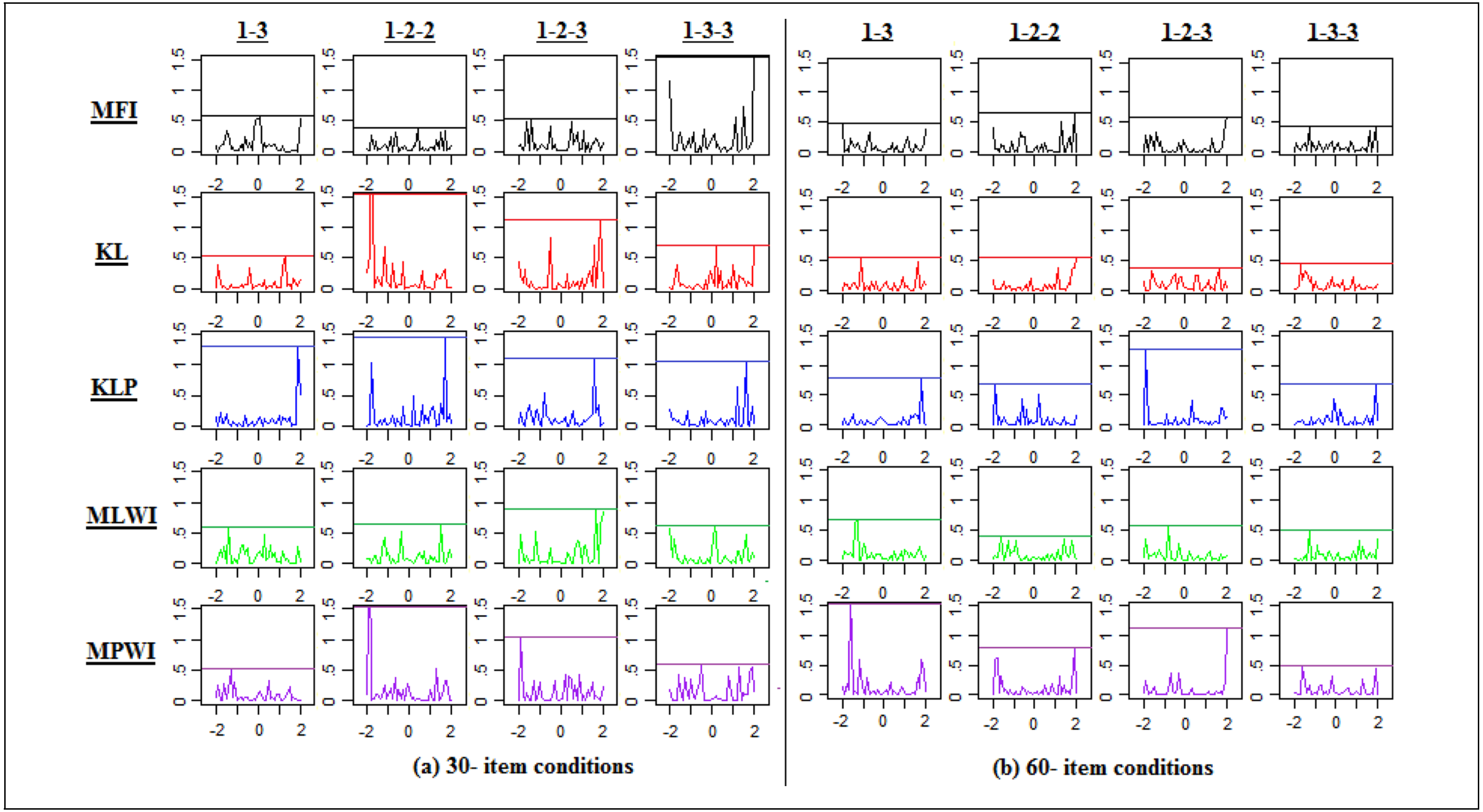


Figure 3. Conditional root mean square errors across all study conditions.

Note. MFI = maximum Fisher information; KL = Kullback–Leibler; KLP = posterior Kullback–Leibler; MLWI = maximum likelihood weighted information; MPWI = maximum posterior weighted information.

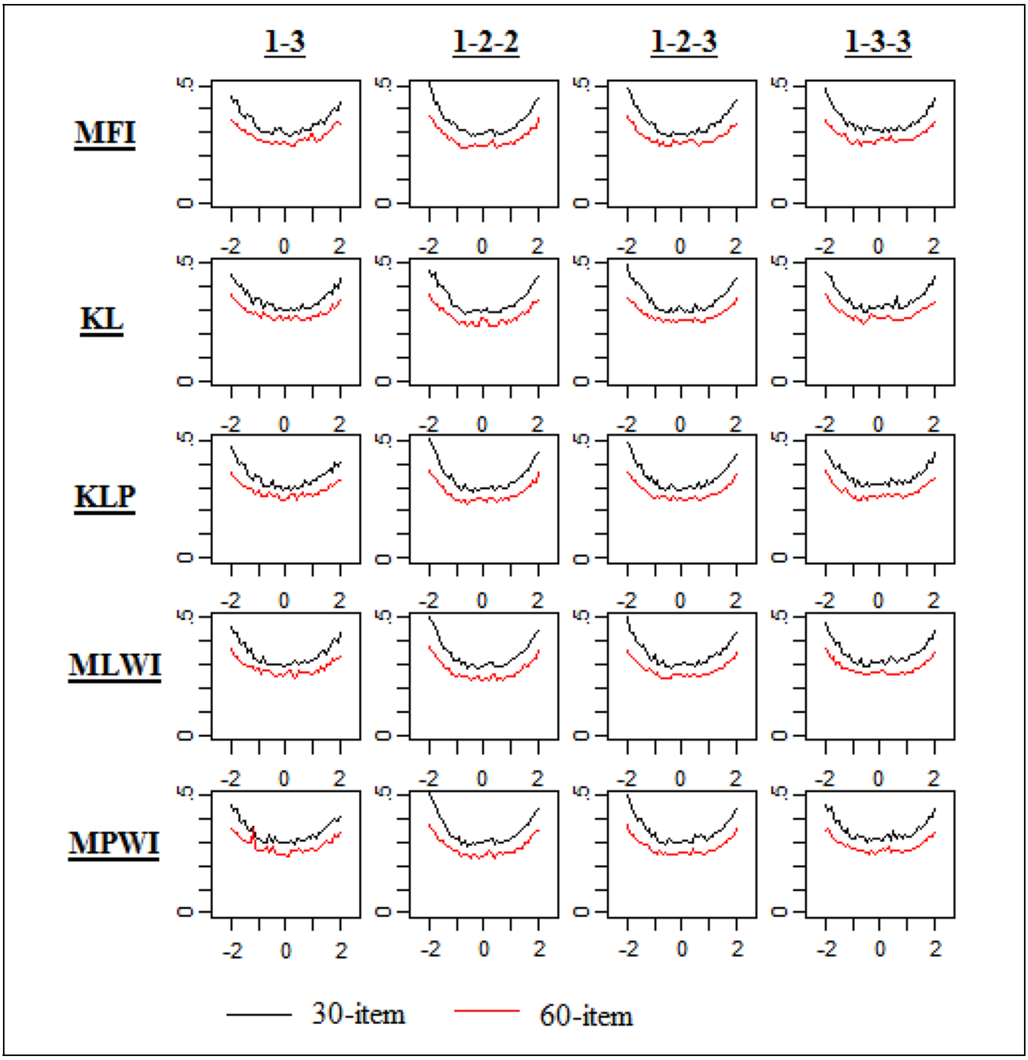


Figure 4. Conditional standard error of measurements across all study conditions.
Note. MFI = maximum Fisher information; KL = Kullback–Leibler; KLP = posterior Kullback–Leibler; MLWI = maximum likelihood weighted information; MPWI = maximum posterior weighted information.

findings. Increasing test length improved the overall outcomes. This was more obvious when CSEMs were investigated (see Figure 4).

On the conditional results, again test length affected the conditional outcomes much more than other study conditions. When the test length was 30, the 1-3 panel design produced more stable conditional absolute biases and RMSEs for each routing method except MFI. Initially it was thought that due to fewer number of adaptation points, the two-stage design (e.g., 1-3 panel design) could have been panelized and could not have a fair chance to compete with three-stage MST designs. However, the type of ca-MST design surprisingly did not affect the study outcomes; in fact, in many cases the two-stage design outperformed three-stage panel design. This was due to number of item in the modules because with all designs having equal total test length, there were more items in both the routing and Stage 2 modules in 1-3 panel design. In addition to the total test length, type of panel design affected the conditional outcomes because having less number of stages allowed placing higher number of items in the modules. Thus, for lower total test lengths (e.g., 30 or less), the 1-3 panel design is recommended with the exception MFI. If the MFI is the interested routing method for a lower test length ca-MST, using the 1-2-2 panel structure is recommended. The routing methods of KL, MLWI, and MPWI are the best choices for the 1-3 30-item conditions. When the test length increased to 60-item, with the exception of MPWI, all routing methods produced better absolute bias. The effect of increasing test length was more obvious for KL method. For long test conditions, 1-3-3 panel design is recommended because all routing methods were more stable under this panel design. Furthermore, for long test ca-MSTs, the routing methods of MFI, KL, and MLWI are the best choices regardless of the type of panel design. However, KLP and MPWI are not recommended in any particular 60-item conditions unless the panel design is 1-3-3. Overall, MLWI is the method of choice due to its consistently low bias and RMSE and should be considered at least as efficient as MFI in most circumstances.

The main criticism with the MFI in CAT framework is that it selects the best items for the provisional thetas but in early stages of CAT, there is a huge discrepancy between provisional and true thetas, therefore, the selected items are actually not the best items (Chang & Ying, 1996). Furthermore, MFI keeps selecting best items again and again, and consumes most informative items across the provisional thetas in early stages of CAT (e.g., most informative items quickly reach the predefined maximum exposure rate). This is why it is not recommended for short test CAT applications (S.-Y. Chen et al., 2000), and that was the motivation of proposing other item selection methods in CAT. However, as there are no such item selection problems in the ca-MST framework, it is as efficient as or better than its modified versions (e.g., MLWI and MPWI).

In this study, some popular item selection methods proposed for use in CAT applications were borrowed. A further study can be conducted by adopting other item selections and using as a routing method in ca-MST framework. In this simulation work, the most common ca-MST panel designs were used. This simulation can also be tested with more complex panel designs (e.g., 1-3-4, 1-4-4, 1-3-4-5, 1-3-5-6-7). This will allow researchers to see the efficiency of these routing methods under varying number of items in the modules with total test being equal. In this study, when assembling the panel designs, some fixed theta points in Stage 2 and Stage 3 modules were selected (e.g., $\theta = 0$, $\theta = -0.5$, $\theta = 0.5$, $\theta = -1$, and $\theta = 1$). A further study can have the panels assembled by maximizing at three to five uniformly spaced theta points instead. Furthermore, researchers should test and compare these methods in both CAT and ca-MST framework with the same item bank. Using the same method in both test administration models will allow one to see what would be lost by administering CAT versus ca-MST when they use the same item bank and item/module selection strategy.

Appendix

Maximum Fisher Information (MFI)

The information for item i at the trait level of θ is calculated as

$$I_i(\theta) = \frac{\left[\frac{\partial P_i(\theta)}{\partial \theta}\right]^2}{P_i(\theta) Q_i(\theta)}.$$

For the 3PL IRT (three-parameter logistic item response theory) model, this equation becomes

$$I_i(\theta) = \frac{2.89a_i^2(1-c_i)}{[c_i + \exp[1.7a_i(\theta-b_i)]][1+\exp[-1.7a_i(\theta-b_i)]]^2},$$

where a_i is the item discrimination parameter, b_i is the item difficulty parameter, c_i is the guessing parameter, $P_i(\theta)$ is the probability of getting item i correct, and $Q_i(\theta)$ is the probability of getting item i incorrect and $Q_i(\theta) = 1 - P_i(\theta)$.

For an estimated trait level of $\hat{\theta}$, MFI selects the next item j among the available items from the item bank that maximizes Fisher information at this estimated trait level by

$$j = \arg\max I_i(\hat{\theta}).$$

Maximum Likelihood Weighted Information (MLWI)

As detailed in the article, MLWI is the product of information function at a fixed theta value and $L(\theta|X_i)$, the likelihood of the fixed theta after i items have been selected. It attempts to find most informative item j over a range of theta values instead of a fixed value by weighting MFI using the theta estimation interval. This results in the equation

$$j = \arg\max \int_{-\infty}^{\infty} I_i(\theta)L(\theta|X_i)\mathrm{d}(\theta).$$

In this study, the range of the integration for this and the following methods was [−4, 4].

Maximum Posterior Weighted Information (MPWI)

MPWI replaces the likelihood weight function by a posterior theta distribution based on previously administrated items and selects the next item that solves

$$j = \arg\max \int_{-\infty}^{\infty} L(\theta|X_i)I_i(\theta)\pi(\theta)\mathrm{d}(\theta),$$

where $L(\theta|X_i)$ is the likelihood function for theta after i items have been selected, $I_i(\theta)$ is the information function, and $\pi(\theta)$ is the prior distribution of theta. For this study, the prior distribution was set to be a standard normal distribution.

Kullback–Leibler (KL)

The KL information is the nonsymmetric distance between two likelihoods at two estimated trait levels (θ, $\hat{\theta}$), denoted $\mathrm{KL}_i(\theta||\hat{\theta})$. The equation to select the next item that maximizes KL information is

$$j = \arg\max \int_{-\infty}^{\infty} \mathrm{KL}_i(\theta||\hat{\theta})L(\theta|X_i)\mathrm{d}(\theta),$$

where $\mathrm{KL}_i(\theta||\hat{\theta})$ is calculated as

$$\mathrm{KL}_i(\theta\|\hat{\theta}) = P_i(\hat{\theta})\ln\left[\frac{P_i(\hat{\theta})}{P_i(\theta)}\right] + \left[1 - P_i(\hat{\theta})\right]\ln\left[\frac{1 - P_i(\hat{\theta})}{1 - P_i(\theta)}\right].$$

Posterior Kullback–Leibler (KLP)

The KLP method weights the current KL information and likelihood function by the prior distribution of theta $\pi(\theta)$, and then selects the next item that maximizes KL information with the equation

$$j = \arg\max \int_{-\infty}^{\infty} \mathrm{KL}_i(\theta\|\hat{\theta})L(\theta|X_i)\pi(\theta)\mathrm{d}(\theta).$$

The prior distribution $\pi(\theta)$ was set as a standard normal distribution.

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.

Supplemental Material

Supplementary material is available for this article online.

References

Armstrong, R. D., Jones, D. H., Li, X., & Wu, I.-L. (1996). A study of a network-flow algorithm and a noncorrecting algorithm for test assembly. *Applied Psychological Measurement*, *20*, 89-98.

Barrada, J. R., Olea, J., Ponsoda, V., & Abad, F. J. (2010). A method for the comparison of item selection rules in computerized adaptive testing. *Applied Psychological Measurement*, *34*, 438-452.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 17-20). Reading, MA: Addison-Wesley.

Bock, D. R., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, *6*, 431-444.

Chang, H.-H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement*, *20*, 213-229.

Chen, P.-H., Chang, H.-H., & Wu, H. (2012). Item selection for the development of parallel forms from an IRT-based seed test using a sampling and classification approach. *Educational and Psychological Measurement*, *72*, 933-953.

Chen, S.-Y., & 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*, 149-174.

Chen, S.-Y., Ankenmann, R. D., & Chang, H.-H. (2000). A comparison of item selection rules at the early stages of computerized adaptive testing. *Applied Psychological Measurement*, *24*, 241-255.

ILOG. (2006). *ILOG CPLEX 10.0* [User's manual]. Paris, France: ILOG S.A.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

Luecht, R. M. (2000, April). *Implementing the computer-adaptive sequential testing (CAST) framework to mass produce high quality computer-adaptive and mastery tests*. Symposium paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Luecht, R. M., Brumfield, T., & Breithaupt, K. (2006). A testlet assembly design for adaptive multistage tests. *Applied Measurement in Education*, *19*, 189-202.

Luecht, R. M., & Nungester, R. J. (1998). Some practical examples of computer-adaptive sequential testing. *Journal of Educational Measurement*, *35*, 229-249.

Luecht, R. M., & Sireci, S. G. (2011). *A review of models for computer-based testing* (Research Report RR-2011-12). New York, NY: The College Board.

Patsula, L. N., & Hambleton, R. K. (1999, April). *A comparative study of ability estimates from computer adaptive testing and multi-stage testing*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Montreal, Quebec.

R Development Core Team. (2009-2015). *R: A language and environment for statistical computing, reference index* (Version 2.2.1). Vienna, Austria: R Foundation for Statistical Computing. Available from http://www.R-project.org

Schnipke, D. L., & Reese, L. M. (1999). *A comparison [of] testlet-based test designs for computerized adaptive testing* (Computerized Testing Report 97-01). Newtown, PA: Law School Admissions Council.

van der Linden, W. J. (1998). Bayesian item selection criteria for adaptive testing. *Psychometrika*, *63*, 201-216. doi:10.1007/BF02294775

van der Linden, W. J., & Glas, C. A. (Eds.). (2000). *Computerized adaptive testing: Theory and practice*. Dordrecht, The Netherlands: Kluwer Academic.

Veerkamp, W. J. J., & Berger, M. P. F. (1997). Some new item selection criteria for adaptive testing. *Journal of Educational and Behavioral Statistics*, *22*, 203-226. doi:10.3102/10769986022002203

Wang, W.-C., & Chen, P. -H. (2004). Implementation and measurement efficiency of multidimensional computerized adaptive testing. *Applied Psychological Measurement*, *28*, 295-316.

Weissman, A., Belov, D. I., & Armstrong, R. D. (2007). *Information-based versus number-correct routing in multistage classification tests* (LSAC Computerized Testing Report No. 07-05). Newtown, PA: Law School Admission Council.

Yan, D., von Davier, A. A., & Lewis, C. (Eds.). (2014). *Computerized multistage testing: Theory and applications*. Boca Raton, FL: CRC Press.

Zenisky, A. L. (2004). *Evaluating the effects of several multi-stage testing design variables on selected psychometric outcomes for certification and licensure assessment* Unpublished doctoral dissertation, University of Massachusetts, Amherst.

Zenisky, A. L., Hambleton, R. K., & Luecht, R. M. (2010). Multistage testing: Issues, designs, and research. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 355-372). New York, NY: Springer.