The Development of MST Test Information for the Prediction of Test Performances

Educational and Psychological
Measurement

2017, Vol. 77(4) 570–586

© The Author(s) 2016

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0013164416662960

journals.sagepub.com/home/epm

**Ryoungsun Park¹, Jiseon Kim²,
Hyewon Chung³, and Barbara G. Dodd⁴**

Abstract

The current study proposes novel methods to predict multistage testing (MST) performance without conducting simulations. This method, called MST test information, is based on analytic derivation of standard errors of ability estimates across theta levels. We compared standard errors derived analytically to the simulation results to demonstrate the validity of the proposed method in both measurement precision and classification accuracy. The results indicate that the MST test information effectively predicted the performance of MST. In addition, the results of the current study highlighted the relationship among the test construction, MST design factors, and MST performance.

Keywords

multistage testing, measurement precision, simulation method, computer-based testing

Introduction

Whether in natural or social sciences, one of the utilities of models (or theories) is explaining things in the real world. Within the context of measuring human traits,

¹Wayne State University, Detroit, MI, USA

²University of Washington, Seattle, WA, USA

³Chungnam National University, Daejeon, South Korea

⁴University of Texas, Austin, TX, USA

Corresponding Authors:

Ryoungsun Park, Educational Evaluation and Research, College of Education, Wayne State University, 5425 Gullen Mall #351, Detroit, MI 48202, USA.

Email: fy3504@wayne.edu

Hyewon Chung, Department of Education, Chungnam National University, 99 Daehak-ro, Yuseong-gu, Daejeon 34134, South Korea.

Email: hyewonchung7@gmail.com

item response theory (IRT) has successfully described in probabilistic terms how examinees respond to items on a test. As a modern measurement theory, IRT conceptualizes the responses to test items through mathematical models using item parameters and the latent variables being measured (Lord & Novick, 1968). Through item and person parameters, IRT can predict the response of an examinee to items even if he or she did not take the item before. This capability of predicting responses to any item allows test designers to develop advanced test forms, such as adaptive tests.

Compared to conventional linear-format tests, the efficiency of the adaptive test is enhanced by maximizing the information for each examinee. Computerized adaptive testing (CAT), often called fully adaptive testing, administers one item at a given time to the individual's ability level (Jodoin, Zenisky, & Hambleton, 2006). Multistage testing (MST), also called computerized sequential adaptive testing, is another variation of adaptive testing that administers a set of items to examinees (Luecht & Nungester, 1998). In MST, test forms are built before the actual administration in the unit of panels and modules (i.e., collection of items) within a panel (Luecht, 2000; Luecht, Brumfield, & Breithaupt, 2006). The administration algorithm routes each examinee from one module to another, while abilities are estimated at the end of each module (Luecht & Nungester, 1998).

In any adaptive testing forms, it is essential to understand the performance of a test before fully implementing it in the real world. Test information function, which is the accumulation of information function for individual items within the test form, provides measurement precision for examinees across ability levels (Embretson & Reise, 2000). In the context of adaptive testing, test information is not readily available because the test forms are not predetermined. Under the CAT, for example, it is highly unlikely for two examinees to receive the same set of items even if their ability levels are similar or equal. For a similar token, examinees are likely to be exposed to different modules under the MST. Therefore, when the need arises to investigate the performance of CAT and MST, some sort of simulation methods have been the de facto choice of methodology.

Three simulation approaches have been popular for CAT simulation, namely, Monte Carlo (MC), post hoc, and hybrid methods, which can be applied to answer different research questions (Nydyck & Weiss, 2009). MC assesses the performance of a CAT algorithm using the samples of items and examinees from known distributions of item parameters and examinees' abilities. Meanwhile, post hoc simulations begin with an actual (i.e., true) response matrix for examinees to a set of items. While MC simulates the responses to each item for generated examinees, the post hoc method uses examinees' true responses to a set of items. One of the applications of the post hoc method is to assess the performance of a specific CAT algorithm for specific examinees taking specific tests. For example, Weiss and Gibbons (2007) assessed the performance of the bifactor CAT algorithm to true scores originally obtained using a paper-and-pencil format for 148 examinees to 626 items. Such a simulation provides answers to questions such as the reduction of test length when

linear tests are converted to the CAT format. The hybrid method allows CAT simulation when complete responses to items are not available (i.e., partial response matrix). This partial response matrix, for example, can occur if not all examinees answer all items due to test security constraints. The hybrid method first fills in the partial response matrix to form a full response matrix, then performs CAT simulation as if the generated full response matrix is a golden one.

The main simulation method for MST, on the contrary, has been a variation of the MC method (conveniently called the ability-level MC method in this article). An ability-level MC method constructs MST using known item parameters, and MC methods are applied to simulate examinees' abilities and their responses to items (Harwell, Stone, Hsu, & Kirisci, 1996). The popularity of the ability-level MC method is attributed to studies on MST that have focused on assessing MST performance for already operational test settings (e.g., American Institute of Certified Public Accountants [AICPA]; National Council Licensure Examination [NCLEX]; Graduate Management Admission Test [GMAT]). These tests were once either paper-pencil tests or CAT format, and MST was evaluated as an alternative approach. The main question for the ability-level MC simulation is querying what will be the MST performance if a large number of examinees were to take the test developed in the MST format. Therefore, simulation studies using real items and simulated examinees have been the predominant choice of methodology by many (Hambleton & Xing, 2006; Jodoin et al., 2006; Luecht, 2003; Park, Kim, Chung, & Dodd, 2014; Patsula, 1999; Patsula & Hambleton, 1999; van der Linden & Reese, 1998; Xing & Hambleton, 2004; Zenisky, 2004; Zheng & Chang, 2015). Recently, a full-MC MST simulation tool was published (Han, 2013). This software program generates three-parameter logistic model parameters and the examinee's abilities under known distributions specified by the user. One immediate benefit of this tool is that it enables researchers to perform an MST simulation without real items. In addition, it can be useful if the research questions are not particularly tied to any operational item pool. However, researchers need to consider the validity of the underlying distributions for items and examinees that they specify.

One of the main goals of a simulation study for a test is to evaluate its measurement precision. In psychometric applications, the measurement precision of a test refers to how accurate the test is in measuring the psychological construct of interest (e.g., ability in educational application). One of the fundamental concepts in IRT involves psychometric information, which has a direct relationship to measurement precision in the test. In IRT, information is a function of the latent trait and item parameters that indicate how accurately an item measures an examinee's ability across the ability levels (De Ayala, 2013; Embretson & Reise, 2000). This item information function describes certain level of measurement precision regardless of whether the ability matches the item's difficulty or not. Another crucial characteristic of item information function is that the information for the entire test (i.e., test information) can be obtained by adding up individual information for each item within the test.

Through this additivity of information function researchers can predict the performance of a test. The standard error of ability estimates is

$$SE(\theta) = \frac{1}{\sqrt{TI(\theta)}} \quad (1)$$

where $TI(\theta)$ is the test information function.

Obtaining test-level information is a challenging task for adaptive test settings because a large number of test forms can be administered to examinees. Thus, a typical methodology for assessing the performance during the test development phase is conducting simulation studies (either MC or ability-level MC). The current study proposes novel methods for deriving not only measurement precision but also classification accuracy for MST without performing actual simulations. First, test-level information for MST (i.e., MST test information) is derived. Using the MST test information, the measurement precision of MST can be calculated based on Equation (1). Second, the MST test information is used to predict the classification accuracy of MST, where the goal of the test is to classify examinees into a set of categories (i.e., pass or fail).

The rest of this article is organized as follows. First, the MST test information is explained in detail regarding its rationale and derivation. Second, the MST test information is extended and applied to calculate the classification accuracy for MST. Third, simulation studies were conducted to demonstrate that the performance of MST is well predicted by the suggested methods. The immediate benefits and broader implications of MST test information are discussed.

Method

MST Test Information

Prior to the full description of MST test information, a couple of crucial observations for MST test forms should be addressed. First, while CAT is considered to be fully adaptive testing, the adaptation in MST occurs at the level of the module, which is a set of items. Once an examinee finishes all the items in a module, a tentative ability is estimated, and a module matching this estimated ability is then administered. Therefore, MST has been known to have characteristics of both the linear test and the fully adaptive test. One characteristic of MST seen from the aspect of a linear test is that test forms are constructed before the test administration according to the test blueprint. Second, modules are typically formed to provide a large amount of information in a certain range on the ability scale, so that they provide high measurement precision for those to whom they are routed. In other words, MST modules target specific difficulties. For instance, in the 1-3-3 MST structure, modules in the second and third stages are constructed to cover namely easy, medium, and hard difficulties. For the detail description of 1-3-3 MST structure, readers should refer to page 231 in Luecht & Nungester's (1998) work. Thus, among all possible pathways, special ones

called primary pathways, in which examinees do not change difficulties at stages beyond the first one (e.g., the second and third stages for 1-3-3 MST), are formed. Each primary pathway can be considered a linear test form sharing a common routing module (i.e., the first module in the MST panel). Third, previous research on MST provides empirical evidence supporting that a large proportion of examinees are routed through the primary pathways. Zenisky (2004) reported the pathway usage frequency of 1-3-3 and 1-2-2 MST structures developed for certification and licensure assessments. For the three routing methods of a defined population interval (DPI; Luecht et al., 2006), proximity, and number-correct scores (NC), the average proportions of examinees following primary pathways were 81.74%, 81.6%, and 71.15% for the 1-3-3 structure and 90.13%, 79.8%, and 81.34% for the 1-2-2 MST design. Furthermore, Kim, Chung, Park and Dodd (2013) compared MST panel designs using the partial credit model and reported that the use of primary pathways was on average 73% and 78% for approximate maximum information (AMI) and DPI routing methods, respectively.

MST Test Information Derivation

The derivation of MST test information involves two steps: (1) test information construction and (2) information averaging. For the first step, information functions are calculated for each primary pathway (i.e., primary pathway information) using the results from the MST test assembly. Information function of each primary pathway is partitioned into two parts: active and inactive. The active part of information is related to the range of theta where information function has an actual contribution to the ability estimation, whereas the inactive part does not contribute to the test performance. For example, the right end (on the ability scale) of the information function for easy primary pathway is active because examinees at the right end (high ability students) are not likely to be routed to easy pathways. The active part of primary pathway information functions are determined such that they will not overlap on the ability scale. In this way, a single test information function is formed when they are placed on top of each other on the ability scale. For example, the MST test information for 1-3-3 MST can be expressed mathematically as

$$TI_{\text{mst}}(\theta) = 1(\theta < C_1)PI_{\text{easy}}(\theta) + 1(C_1 \leq \theta \leq C_2)PI_{\text{mid}}(\theta) + 1(\theta > C_2)PI_{\text{hard}}(\theta) \quad (2)$$

where $1(x)$ is a function equal to 1 if the condition x is true and 0 otherwise, $PI_y(\theta)$ is the pathway information function for a corresponding difficulty y , and C_1 and C_2 are boundaries on the theta scale separating active and inactive parts. For example, for the 1-3-3 design, C_1 and C_2 may be set to the midpoint (intersection) between easy and medium difficulty pathways as well as the midpoint of medium and hard difficulty pathways, respectively. If actual test construction is achieved using the 1-3-3 design so that the peak locations of pathway information functions are at the desired locations, the amount of information at the peak is almost equal, and pathway

Figure 1. MST test information function formed from three primary pathway information functions.

information functions are symmetric, then the intersection between any two pathway information functions will be located at the midpoint of two peaks. In reality, however, we cannot expect to achieve ideal test constructions according to the specification in the target test information functions. Therefore, it is reasonable to take the intersection of two pathway information functions as the boundary, which reflects actual test administration that routes examinees to pathways of larger information for high measurement precision.

Figure 1 shows an example of the 1-3-3 MST test information function constructed from three pathway information functions. The peaks of the pathway information functions for the easy, medium, and hard difficulty are located at approximately $-0.8, 0.0,$ and 1.2 , respectively, on a theta scale. The boundaries between pathways are at $C_1 = -0.6$ and $C_2 = 0.6$. The MST test information function is formed by concatenating all active parts of pathway information functions.

Information Averaging

The MST test information in Figure 1 shows sharp edges at the boundaries where two pathway information functions cross. As examinees at or around boundaries (i.e., C_1 and C_2) are more likely to be routed through secondary pathways, one can anticipate a blending (or mixture) of information from both pathways. For example, an examinee with a theta of around -0.6 may take the easy pathway up to the second stage, then moves to a medium difficulty in the last stage of the 1-3-3 MST structure depending on how well he/she performs in the second stage. To incorporate the secondary pathway routings into the calculation, the simple moving averages are used. Assuming that the ability scale is divided into a finite number of quadrature points, the simple moving average of 5 points, for example, is expressed as

$$Y(\theta) = \frac{\sum_{i=1}^5 X(\theta + i - 3)}{5} \quad (3)$$

where $Y(\theta)$ is the moving average of $X(\theta)$ for 5 points around the theta. This averaged MST test information function is then used to calculate the (analytic) standard errors to predict the MST performance in terms of its measurement precision according to Equation (1).

Predicting MST Classification Accuracy

When MST is applied to licensure or certification tests, one of the primary interests is its classification accuracy. Classification accuracy is defined as the proportion of examinees whose classification outcomes from a test match theoretical results based on their true abilities. Classification error, on the other hand, is the proportion of examinees whose outcome do not match their true categories. The classification error in pass/fail test is often expressed in terms of false positive error rates (FPER) and false negative error rates (FNER). FPER is defined as the proportion of examinees who pass the test even if true ability is below the cutoff score (i.e., falsely passing the test), whereas FNER is the proportion of examinees who fail to pass the exam even though their true ability is above the cutoff score (i.e., falsely failing the test).

The fundamental idea for predicting the classification error for MST is from the observation that standard error calculated from MST information function provides the quantification on the uncertainty (or accuracy) of the ability estimates at a given theta. Assume that the ability estimate (i.e., a random variable θ_1) follows a normal distribution of $N(\theta_1, SE(\theta_1))$, and the cutoff score of the test (i.e., θ_c) is higher than θ_1 (i.e., $\theta_1 < \theta_c$). Then, the proportion of examinees beyond the θ_c indicates the FPER at θ_1 . This conditional FPER at θ_1 can be calculated as $1 - \Phi(\theta_c)$, where $\Phi(\theta_c)$ is the cumulative distribution function defined as $P(\theta < \theta_c)$. By the same token, if θ_c is lower than θ_1 , the FNER is obtained from $\Phi(\theta_c)$. As the population of examinees is typically assumed to follow a certain distribution, the total amount of the error rates needs to consider the relative frequency of examinees given theta. To calculate total

error rates, conditional error rates need to be weighted by the relative proportion of examinees on the ability scale.

The actual calculation of total error rates can be easily approximated using a numerical method called quadrature points. Quadrature points divides the continuous ability scale into a finite number of intervals, in which a single standard error value may be representative for each interval. The total error rates are then simply the summation of the error rates for all quadrature points. Even though the entire theta scale needs to be considered for the calculation for the error rates, sufficient accuracy can be achieved by including a handful quadrature points around the cutoff score. This is because examinees around the cutoff score are major contributions to the actual classification errors. Based on a finite number of quadrature points, mathematical representations for FPER and FNER are

$$\text{FPER} \cong \sum_{i=1}^I w_i f(\theta_i) (1 - \Phi(z_i)) \quad (4)$$

$$\text{FNER} \cong \sum_{i=1}^I w_i f(\theta_i) \Phi(z_i) \quad (5)$$

where I is the number of the quadrature points, w_i is the width of the i th quadrature point (typically a constant value indicating equal width across quadrature points), $f(\theta_i)$ is the y coordinate of the standardized normal distribution on θ_i , $\Phi(z_i)$ is the cumulative distribution function of the standardized normal distribution, and z_i is the standardized distance between the current θ_i and the cutoff score θ_c , which is expressed as

$$z_i = \frac{\theta_c - \theta_i}{SE(\theta_i)} \quad (6)$$

For a concrete illustration of classification error calculation, an example is presented. Let us assume that the MST test is administered to classify examinees into two categories (pass or fail), with the cutoff score of $\theta_c = 0.0$ and MST test information and its standard errors were calculated according to Equations (2) and (1), respectively. Figure 2 shows three theta intervals centered at quadrature points of $\theta_i = -0.1, -0.3$, and -0.5 (thus, $w_i = 0.2$.) for the estimation of FPER.

Table 1 indicates the FPER calculations. Hypothetical $SE(\theta_i)$ values are assumed and the test information around the cutoff score is approximately 12, resulting in standard error values about 0.29. Values for $f(\theta_i)$ indicate the probability density at θ_i for the standardized normal distribution. The total FPER is 0.043 (i.e., the sum of $FPER_i$ from the table), implying that 4.3% of examinees are expected to falsely pass this MST test. Furthermore, note that as theta moves away from the cutoff score, the relative contribution to the total error rates decreases rapidly.

Figure 2. Quadrature points and their interval for the classification error calculation assuming a normal population distribution.

Table 1. Calculation Example of FPER.

i	θ_i	θ_c	$SE(\theta_i)$	z_i	$1 - \Phi(z_i)$	$f(\theta_i)$	w_i	$FPER_i$
1	-0.1	0	0.289	0.345	0.365	0.397	0.200	0.029
2	-0.3	0	0.290	1.034	0.151	0.381	0.200	0.011
3	-0.5	0	0.289	1.727	0.042	0.352	0.200	0.003

Note. $FPER_i$ represents the error rates for each i th interval and FPER is 0.043, which is the sum of $FPER_i$.

Simulation Studies

To demonstrate that the proposed methods are able to predict measurement precision and classification accuracy in the MST context, two simulation studies were performed using mixed-format items. The mixed-format item pool used in the current study is the science subset of the 1996 National Assessment of Educational Progress (NAEP), which was calibrated according to the generalized partial credit model (GPCM; Muraki, 1992). MST test forms were constructed using linear programming method (LP; van der Linden, 2005). For the ability estimation, expected a priori (EAP; Bock & Mislevy, 1982) method was used with a standardized normal prior.

Conditions for the First Simulation (Measurement Precision)

MST forms were constructed while considering four design factors: two levels of test length, two MST structures, two routing module designs, and four proportions between dichotomous and polytomous items. These MST design conditions are chosen so that the validity of MST test information method can be generalized to a wide range of realistic mixed-format test settings (Park, 2015). For the test length condition, 40 and 60 total points were considered in the study. A test length of 60 points has been recommended for licensure and diagnostic tests in the MST context (Jiao, 2003) while 40 points were included to investigate the impact of shorter tests. The two MST structures included in the study were 1-3-3 and 1-2-2 structures. MST of 1-3-3 forms are commonly used for the ability estimation and 1-2-2 forms are widely accepted for licensure and certification tests (Jodoin et al., 2006; Zenisky, 2004). Two routing module designs varied the proportion of polytomous and dichotomous items within the routing module in MST: purely dichotomous and a mixture of dichotomous and polytomous items. Finally, the study varied the proportion of polytomous items within the MST design from small to moderately large proportions (i.e., 10%, 30%, 50%, and 70%). The proportion of dichotomous and polytomous items in mixed-format tests is an important factor that to date has not received much attention (Kim & Dodd, 2014).

In summary, the total number of study conditions is 32: 4 (proportion of polytomous items) by 2 (total points) by 2 (MST structures) by 2 (routing module designs). In each simulation condition, 1,000 examinees from a standard normal distribution were generated and 100 replications were carried out; the standard errors of estimation were averaged over the 100 replications. During the simulation, the AMI routing method was used to assign examinees to modules likely to provide maximum information regarding examinees' abilities. The empirical standard errors obtained from the simulations were compared to the standard errors calculated from the MST test information.

Conditions for the Second Simulation (Classification Accuracy)

The MST simulation was performed to estimate the classification accuracy. The 1-3-3 MST form was constructed using the LP method while content balancing constraints were considered (van der Linden, 2005). Then, estimated thetas from the MST simulation were used to determine if examinees were classified into pass or fail according to three cutoff scores -0.524 , 0.0 , and 0.524 . They correspond to 30%, 50%, and 70% passing rates, respectively, assuming that examinees are normally distributed. Two classification errors (i.e., FPER and FNER) and the total error rate (TER)—namely, the sum of FPER and FNER—were calculated. As in the first simulation, 100 replications were performed while 1,000 examinees from a standard normal distribution were generated for each replication.

Figure 3. Grand mean standard errors averaged across 32 study conditions.

Results

Simulation results showed that, on average, 89% of examinees were routed through the primary pathways, strengthening the observations made in previous research. Figure 3 compares analytic and empirical grand mean standard errors averaged across 32 study conditions, presenting overall match of results from two methodologies. Across a wide range of ability levels, analytic standard error values were similar to those obtained empirically. Overall, the analytic *SEs* are very close the empirical *SEs*. The discrepancy of *SEs* in the lower end of theta is caused by the inaccuracy of estimation in simulation studies, which is mainly due to the small number of examinees in a normally distribution and the lack of highly informative items in the pool.

Table 2 shows the differences in analytic and empirical standard errors for each simulation condition. The two methods produce MST performance values very close to each other. However, at lower ability levels, analytic standard errors tend to be greater than empirical values, as indicated by consistently positive difference values.

Table 3 presents the classification accuracy and error rates obtained from both simulation results and analytic calculations across three passing rates (30%, 50%, and 70%). The two methods produced very close results in terms of FNER and FPER and, consequently, TER and CCR values. For the 70% passing rate condition, the two approaches showed slightly larger FNER compared to FPER, while the FNER is smaller than FPER for the 30% passing rate. For the 50% passing rate, both FNER and FPER values are close to each other. Thus, two methods produced virtually same

Table 2. Differences Between Analytic and Empirical Standard Errors.

Design factors			Theta points																
PPI	TL	RMD	MS	-3.5	-3.0	-2.5	-2.0	-1.5	-1.0	-0.5	0.0	0.5	1.0	1.5	2.0	2.5	3.0	3.5	
10	40	Dichotomous		1-2-2	0.225	-0.006	-0.078	-0.053	-0.014	-0.013	0.009	-0.010	-0.021	-0.048	-0.042	0.001	-0.064	-0.062	
				1-3-3	0.120	-0.021	-0.050	-0.012	-0.016	-0.005	-0.001	-0.013	0.012	-0.009	-0.022	-0.024	-0.018	-0.040	-0.077
	Mixed		1-2-2	0.192	-0.010	-0.101	-0.021	-0.017	-0.003	-0.012	0.008	-0.006	-0.021	-0.050	-0.035	-0.009	-0.057	-0.061	
			1-3-3	0.061	-0.050	-0.056	-0.018	-0.011	0.002	-0.007	-0.014	0.009	-0.008	-0.019	-0.023	-0.028	-0.077	-0.059	
	60	Dichotomous		1-2-2	0.062	-0.062	-0.042	-0.015	-0.006	-0.003	-0.007	0.006	-0.007	-0.009	-0.028	-0.048	-0.034	-0.060	-0.046
				1-3-3	0.030	-0.020	-0.041	-0.005	-0.009	-0.003	-0.004	-0.011	0.007	-0.009	-0.015	-0.015	-0.014	-0.038	-0.057
30	40	Mixed		1-2-2	0.048	-0.017	-0.047	-0.016	-0.008	0.002	-0.011	0.005	-0.004	-0.008	-0.033	-0.048	-0.029	-0.026	-0.056
				1-3-3	0.007	-0.020	-0.029	-0.005	-0.004	0.000	-0.005	-0.010	0.007	-0.007	-0.012	-0.020	-0.016	-0.059	-0.081
	Dichotomous		1-2-2	0.146	-0.001	-0.072	-0.032	-0.019	-0.009	-0.009	0.009	-0.011	-0.021	-0.053	-0.074	-0.046	-0.040	-0.035	
			1-3-3	0.208	0.010	-0.055	-0.029	-0.030	-0.012	-0.007	-0.016	0.009	-0.011	-0.024	-0.025	-0.017	-0.030	-0.028	
	Mixed		1-2-2	0.264	0.048	-0.081	-0.030	-0.024	-0.007	-0.009	0.008	-0.008	-0.009	-0.018	-0.040	-0.045	-0.036	-0.027	-0.066
			1-3-3	0.042	-0.074	-0.075	-0.019	-0.014	-0.003	-0.009	-0.021	0.008	-0.013	-0.020	-0.039	-0.055	-0.087	-0.017	
50	60	Dichotomous		1-2-2	0.030	-0.063	-0.037	-0.020	-0.012	-0.001	-0.009	0.007	-0.009	-0.010	-0.043	-0.068	-0.071	-0.146	-0.101
				1-3-3	0.016	0.009	-0.032	0.001	-0.014	-0.005	-0.001	-0.009	0.008	-0.009	-0.010	-0.012	0.001	-0.037	-0.038
	Mixed		1-2-2	0.053	-0.031	-0.043	-0.014	-0.009	-0.001	-0.010	0.006	-0.009	-0.010	-0.040	-0.069	-0.087	-0.114	-0.090	-0.090
			1-3-3	0.029	-0.046	-0.027	-0.007	-0.008	-0.003	-0.006	-0.011	0.004	-0.007	-0.013	-0.008	0.001	-0.030	-0.040	
	40	Dichotomous		1-2-2	0.219	0.021	-0.092	-0.066	-0.031	-0.012	-0.016	0.010	-0.009	-0.023	-0.048	-0.040	0.014	0.025	0.064
				1-3-3	0.163	0.028	-0.072	-0.017	-0.031	-0.008	-0.007	-0.015	0.012	-0.009	-0.025	-0.031	-0.024	-0.036	-0.009
60	Mixed		1-2-2	0.223	0.050	-0.052	-0.037	-0.030	-0.007	-0.011	0.009	-0.009	-0.020	-0.051	-0.055	-0.029	-0.012	0.026	
			1-3-3	0.148	-0.049	-0.080	-0.031	-0.023	-0.006	-0.006	-0.015	0.011	-0.005	-0.019	-0.028	-0.013	-0.088	0.053	

(continued)

Table 2. (continued)

Design factors			Theta points																
PPI	TL	RMD	MS	-3.5	-3.0	-2.5	-2.0	-1.5	-1.0	-0.5	0.0	0.5	1.0	1.5	2.0	2.5	3.0	3.5	
60	Dichotomous		I-2-2	0.099	-0.001	-0.072	-0.026	-0.011	-0.006	-0.009	0.008	-0.006	-0.013	-0.037	-0.031	-0.004	-0.003	0.087	
			I-3-3	0.010	-0.030	-0.047	-0.011	-0.010	-0.008	0.000	-0.008	0.008	-0.008	-0.011	-0.007	-0.006	-0.018	-0.034	
		Mixed	I-2-2	0.164	0.004	-0.060	-0.015	-0.019	0.000	-0.009	0.007	-0.007	-0.011	-0.031	-0.025	0.007	-0.007	-0.006	
70	Dichotomous		I-3-3	0.093	-0.045	-0.046	-0.009	-0.015	-0.005	-0.002	-0.013	0.006	-0.011	-0.017	-0.016	-0.003	-0.014	-0.009	
			I-2-2	0.226	0.051	-0.090	-0.050	-0.042	-0.013	-0.016	0.009	-0.008	-0.021	-0.062	-0.020	0.000	0.000	0.000	
		Mixed	I-3-3	0.316	0.081	-0.053	-0.047	-0.037	-0.011	-0.008	-0.011	0.009	-0.009	-0.017	0.005	0.000	0.000	0.000	
60	Dichotomous		I-2-2	0.227	0.001	-0.087	-0.060	-0.047	-0.009	-0.011	0.008	-0.011	-0.021	-0.052	0.000	0.000	0.000	0.000	
			I-3-3	0.205	0.018	-0.086	-0.039	-0.034	-0.008	-0.010	-0.012	0.013	-0.006	-0.016	-0.005	0.000	0.000	0.000	
		Mixed	I-2-2	0.100	-0.026	-0.063	-0.033	-0.019	-0.008	-0.014	0.005	-0.005	-0.010	-0.029	-0.018	-0.036	-0.021	0.004	
	Dichotomous		I-3-3	0.136	-0.006	-0.058	-0.026	-0.028	-0.004	-0.011	-0.007	0.007	-0.007	-0.008	-0.004	0.008	-0.002	-0.075	
			I-2-2	0.159	-0.033	-0.065	-0.021	-0.017	-0.004	-0.008	0.008	-0.006	-0.012	-0.032	-0.017	-0.001	0.012	-0.006	
		Mixed	I-3-3	0.054	-0.050	-0.051	-0.007	-0.014	-0.003	-0.006	-0.012	0.006	-0.008	-0.016	-0.014	0.002	-0.006	0.044	

Note. MST = multistage testing; PPI = proportion of polytomous item; TL = test length; RMD = routing module design; MS = MST structure. SEs were calculated on 13 data points of theta, ranging from -3.0 to 3.0 in increments of 0.5, while the actual theta ranges were from -3.25 to 3.25. For example, the SE values between -3.25 and -2.75 were averaged to represent the measurement accuracy on the theta of -3.0.

Table 3. Comparison of Classification Accuracies Between the Simulation Results and Analytic Estimates.

Passing rate (%)	Simulation results				Analytic results			
	CCR	FNER	FPER	TER	CCR	FNER	FPER	TER
30	91.95	3.72	4.33	8.05	92.10	3.83	4.07	7.90
50	91.18	4.54	4.28	8.82	91.20	4.45	4.35	8.80
70	91.89	4.40	3.71	8.11	91.83	4.28	3.89	8.17

Note. CCR = correct classification rate; FNER = false negative error rate; FPER = false positive error rate; TER = total error rate. All rates are represented in percentages.

results in terms of overall error rates (thus accuracy rates) across three levels of passing rate conditions.

Discussion

The current study suggested novel methods to obtain performance estimates in MST. The analytic derivation of MST test information, which is the estimation of the test information function under MST design, is the essential building block of the new methods. Based on the MST test information, the measurement accuracy of MST can be derived through the concept of the standard error of estimates; it can also be applied to estimate classification accuracy through the quadrature points on the ability scale.

From the methodological perspective, the findings in this study defy the norm. MC simulation has been a de facto method for obtaining MST performance matrices in many studies. The results from the current study imply that researchers now have an additional methodological choice when it comes to MST performance measures. Setting up a simulation environment may take a significant amount of time and effort. Thus, the turnaround time from a design change to simulation results can be quite large depending on the computing power and resources. The proposed method has the potential to remove these steps, and simple spreadsheet calculations are sufficient for deriving MST performances.

In addition to the practical advantages, such as the reduction of simulation time and resources, this methodology is well aligned with the recommendation from *Psychometrika*'s publication policy (Psychometric Society, 1979). The suggestion states that the research problems should be solved using MC methods only if an acceptable solution cannot be obtained using an analytic method. In addition, the results from the traditional MC simulation method at the extreme range of the ability scale may raise questions about its accuracy due to the small sample sizes of examinees (assuming normal distribution of examinees) and low variability in the response patterns. Analytic results do not suffer from this shortcoming as they do not rely on

generated response patterns from small number of samples; thus, they can be considered to be more accurate than the MC simulation results.

Although the immediate benefit of this method is the reduction (or potential elimination) of simulation and researchers are able to obtain performance metrics right after the test construction, the implications are beyond this low-hanging fruit. In his recent publication, Chang (2015) pointed out that a large body of MST-related publications in journals and conferences have not made significant advancements since Luecht and Nungester's (1998) original work. In the article, he suggested a couple of directions, such as developing an innovative MST assembly methodology and exploring new test forms combining MST and CAT (i.e., on-the-fly MST). The findings in this study may provide useful insights regarding the lack of innovation in MST studies. It was demonstrated that two crucial performance measures of MST can be directly calculated from the result of the test assembly, highlighting the strong relationship between test assembly and MST performance beyond and above any other factors. This strong connection between test assembly and MST performance may partially explain why researchers were not able to demonstrate the impact of various other design factors. Using terminologies borrowed from statistics, the outcome of the test assembly procedure is a strong mediating variable accounting for a large portion of the variability in the MST test performance. For example, the impact of the test length to the measurement precision can vary depending on how target information was set during the test assembly rather than the condition of the test length itself. Admittedly, it is highly likely that increasing the test length will increase MST test information. However, this causal impact of test length to the increased measurement precision is not a natural law that always occurs, but rather one of the test assembly specifications that may not be translated to the measurement precision. For this reason, two researchers may draw disparate conclusions on the impact of a certain design factor even if all the other conditions are equal (e.g., item pool, test length, the number of panels) when their test assembly procedures do not match. The findings in the study will contribute to the deeper understanding of the relationship between the test assembly and test performance in the context of MST.

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

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

- Chang, H. H. (2015). Psychometrics behind computerized adaptive testing. *Psychometrika*, 80, 1-20.
- De Ayala, R. J. (2013). *Theory and practice of item response theory*. New York, NY: Guilford.
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- Hambleton, R. K., & Xing, D. (2006). Optimal and nonoptimal computer-based test designs for making pass-fail decisions. *Applied Measurement in Education*, 19, 221-239.
- Han, K. T. (2013). MSTGen simulated data generator for multistage testing. *Applied Psychological Measurement*, 37, 666-668.
- Harwell, M., Stone, C. A., Hsu, T. C., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement*, 20, 101-125.
- Jiao, H. (2003). *The effects of model specification error in item response theory-based computerized classification test using sequential probability ratio test* (Unpublished doctoral dissertation). The Florida State University, Tallahassee.
- Jodoin, M. G., Zenisky, A. L., & Hambleton, R. K. (2006). Comparison of the psychometric properties of several computer-based test designs for credentialing exams with multiple purposes. *Applied Measurement in Education*, 19, 203-220.
- Kim, J., Chung, H., Park, R., & Dodd, B. G. (2013). A comparison of panel designs with routing methods in the multistage test with the partial credit model. *Behavior Research Methods*, 45(4), 1087-1098.
- Kim, J., & Dodd, B. G. (2014). Mixed-format multistage tests: issues and methods. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing: Theory and applications* (pp. 55-67). Boca Raton, FL: CRC Press.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Boston, MA: Addison-Wesley.
- Luecht, R. M. (2000, April). *Implementing the computer-adaptive sequential testing (CAST) framework to mass produce high quality computer-adaptive and mastery tests*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.
- Luecht, R. M. (2003, April). *Exposure control using adaptive multi-stage item bundles*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.
- 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.
- Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16, 159-176.
- Nydicke, S. W., & Weiss, D. J. (2009). A hybrid simulation procedure for the development of CATs. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing*. Retrieved from www.psych.umn.edu/psylabs/CATCentral
- Park, R. (2015). *Investigating the impact of a mixed-format item pool on optimal test designs for multistage testing* (Unpublished doctoral dissertation). The University of Texas, Austin.
- Park, R., Kim, J., Chung, H., & Dodd, B. G. (2014). Enhancing pool utilization in constructing the multistage test using mixed-format tests. *Applied Psychological Measurement*, 38, 268-280.

- Patsula, L. N. (1999). *A comparison of computerized-adaptive testing and multi-stage testing* (Unpublished doctoral dissertation). University of Massachusetts at Amherst.
- Patsula, L. N., & Hambleton, R. K. (1999, April). *A comparative study of ability estimates obtained from computer-adaptive and multi-stage testing*. Paper presented at the meeting of the National Council on Measurement in Education, Montreal, Quebec, Canada.
- Psychometric Society. (1979). Publication policy regarding Monte Carlo studies. *Psychometrika*, 44, 133-134.
- van der Linden, W. J. (2005). *Linear models for optimal test design*. New York, NY: Springer.
- van der Linden, W. J., & Reese, L. M. (1998). A model for optimal constrained adaptive testing. *Applied Psychological Measurement*, 22, 259-270.
- Weiss, D. J., & Gibbons, R. D. (2007). Computerized adaptive testing with the bifactor model. In *Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing*. Retrieved from <http://publicdocs.iacat.org/cat2010/cat07weiss&gibbons.pdf>
- Xing, D., & Hambleton, R. K. (2004). Impact of test design, item quality, and item bank size on the psychometric properties of computer-based credentialing examinations. *Educational and Psychological Measurement*, 64, 5-21. doi:10.1177/0013164403258393
- 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 at Amherst.
- Zheng, Y., & Chang, H. H. (2015). On-the-fly assembled multistage adaptive testing. *Applied Psychological Measurement*, 39, 104-118.