A comparison of panel designs with routing methods in the multistage test with the partial credit model

Jiseon Kim · Hyewon Chung ·
Ryoungsun Park · Barbara G. Dodd

Published online: 7 March 2013
© Psychonomic Society, Inc. 2013

Abstract In this study, we compared panel designs applied with various routing methods in the multistage test (MST) based on the partial credit model in the context of classification testing. Simulations were performed to compare three routing methods and four panel structures. Conditions of two test lengths and three passing rates were also included. The results showed that, regardless of the routing method used, the same panel structure performed similarly in terms of the precision of the classification decision with the same test length condition. The longer test length produced higher accuracy, whereas the 50 % passing rate yielded the lowest accuracy. Finally, all MST conditions performed well in terms of test security.

Keywords Multistage test · Test construction · Automated test assembly

As one approach to computer-based testing, the multistage test (MST) requires that examiners choose and administer sets of items and adapt them to each examinee. Thus, MST is often regarded as a compromise approach between fully computerized adaptive testing (CAT) and paper-and-pencil tests (Jodoin, Zenisky, & Hambleton, 2006). Studies have shown that the MST requires a shorter test length and can provide equal or higher predictive and concurrent validities

relative to linear tests, including paper-and-pencil tests (Betz & Weiss, 1974; Kim & Plake, 1993; Linn, Rock, & Cleary, 1969; Patsula & Hambleton, 1999; Wainer, 1995).

Unlike CAT, which applies a variable test length condition according to the examinees' estimated abilities, MST is a fixed-length test and has several additional elements beyond CAT, comprising panels, modules, stages, and pathways. Panels can be considered to be much like a test form in a conventional test setting, and many panels can be constructed on the basis of the capacity of the item pool and test characteristics. Panels are composed of a specified number of stages, within which several modules are gathered together. Modules are the smallest units, which include sets of items that comprise the different levels of difficulty (e.g., easy, medium, or hard modules). Finally, pathways are considered the routes that examinees can take from module to module within a particular panel, using various routing methods based on estimates of the examinees' abilities (Luecht, 2000).

During the test development process, human review is often a crucial requirement in order to assure test quality (Luecht & Nungester, 1998). Test developers construct MST forms (e.g., panels) prior to the test being administered. Consequently, this gives MST significant advantages as compared to CAT. For CAT, the test form for each examinee is built as the test is being administered. Thus, content specialists cannot review every test form before the test is given to verify that all content requirements are satisfied. Generally, CAT is not capable of incorporating all of the content specifications of the test blueprint, even with the most refined content-balancing algorithm (Keng, 2008). MST thus features a better quality control procedure, because it ensures that content is balanced equally for each examinee before the test is administered.

In addition, MST allows examinees to review previous items within a module, because his or her ability is estimated after finishing an entire module (Patsula, 1999). MST has been researched in the contexts of achievement testing and

J. Kim (✉)
Department of Rehabilitation Medicine, University of Washington,
Seattle, WA, USA
e-mail: jiseonk@u.washington.edu

H. Chung (✉)
Department of Education, Chungnam National University,
Daejeon, South Korea
e-mail: hyewonchung7@gmail.com

R. Park · B. G. Dodd
Department of Educational Psychology, University of Texas,
Austin, TX, USA

classification test settings, such as the Uniform Certified Public Accountant Examination or the United States Medical Licensing Examination (Breithaupt & Hare, 2007). The results of these previous research studies have indicated that MST might be a useful approach for these contexts.

Generally, assembling MSTs has been conducted using computer software such as an automated test assembly (ATA) program. Such ATA programs can produce fixed-length parallel tests with nonstatistical and statistical constraints before the test is administered. ATA can also apply optimization algorithms such as linear programming (LP), heuristics, or both, to create multiple panels, stages, and modules simultaneously in MST (Luecht, Brumfield, & Breithaupt, 2006; Luecht & Nungester, 1998). Whether based on an ATA program or manual test construction, many MST studies have constructed and applied the 1–3–3 MST panel structure. This approach has one module at the first stage and three modules at each of the second and third stages of the test (see, e.g., Davis & Dodd, 2003; Keng, 2008; Kim, Chung, Park, & Dodd, 2011).

During test administration, examinees are assigned to the first stage of one of the multiple panels within the test. After finishing an entire stage, an examinee's ability is then estimated. To estimate ability, the number-correct (NC) scoring method, item response theory (IRT) ability estimation methods such as maximum likelihood estimation (MLE), or expected a posteriori (EAP) estimates are commonly used to route examinees to modules with a different level of difficulty in the next stages of the test (Hendrickson, 2007; Luecht, 2000). The MLE procedure has been used in MST studies by Davis and Dodd (2003) and Kim et al. (2011). Jodoin (2003) and Keng (2008) applied EAP in their MST studies. NC scoring is likely the simplest method that could be used to assign examinees to a module of a subsequent stage. Using NC scoring to estimate each examinee's final ability (at the end of the entire test), however, is not recommended for MST, because the items that each examinee receives are not statistically equivalent (Lord, 1980).

According to the examinee's performance in the current stage, he or she will be assigned (or routed) to one of the modules residing in a subsequent stage. For the present study, we used routing methods (described next) that considered the proportion of examinees within the module or the stage or the relative difficulty level of the subsequent stage modules, given the estimated abilities of examinees. These routing methods have been the ones most frequently used in recent research.

The defined population intervals (DPI) method routes a proportion of examinees to different modules of the next stage (Zenisky, 2004). For example, after finishing their current stage, examinees across modules are rank-ordered according to their estimated abilities. Then, a certain proportion of these examinees are assigned to each module of

the subsequent stage (Jodoin, 2003; Xing, 2001; Zenisky, 2004). In addition to the regular DPI method, examinees can be rank-ordered according to ability estimates within the module and routed to a subsequent-stage module. In the present study, this method is called *module-level DPI* (ML-DPI). The original method of DPI is called *stage-level DPI* (SL-DPI). The DPI methods are relatively easy to implement and provide exposure rates in advance of testing (e.g., Jodoin, 2003; Xing, 2001; Zenisky, 2004). Proportionally assigning examinees to different next-stage modules, however, might be convenient for controlling exposure control rates, but it might not do a good job of properly matching the modules' levels of difficulty to the examinees.

Unlike the DPI methods, the approximate maximum information method (AMI; Luecht et al., 2006) uses information functions to route examinees by determining the intersection between the cumulative test information functions (TIFs) of adjacent modules. Examinees are routed to the next stage that will provide the maximum information of cumulative TIFs, given the examinee's current ability estimates. As an alternative to AMI, some MST studies (e.g., Davis & Dodd, 2003; Keng, 2008; Kim et al., 2011) have selected the next module that will provide maximum information on the basis of the examinees' provisional ability estimates using module TIFs. In the present study, we call this method the *modified version of AMI* (M-AMI; Keng, 2008). As compared to the DPI methods, the advantage of M-AMI is that it uses the available IRT information, just as CAT chooses items. By assigning examinees in this way, however, the distribution of examinees to different modules within each stage will vary. Consequently, some modules will likely be exposed to examinees more often than will other modules. For example, if the examinees' abilities are normally distributed, more examinees will likely be routed into medium-difficulty modules (Zenisky, 2004).

Properly constructing and administering the MST is crucial for producing accurate results, defined as classifying examinees into correct categories (Zenisky, 2004). Most MST studies have investigated panel design (or structures) and routing methods separately. For example, some studies have investigated the performance of variations among panel structures. Jodoin et al. (2006) compared a 1–3–3 panel structure, which included one module for the first stage and three modules for the second and third stages with 60 items. Jodoin et al. also examined a 1–3 panel structure with 40 items using the three-parameter logistic IRT model. The 1–3 panel design showed a slightly lower classification accuracy than did the 1–3–3 panel structure. In addition, Kim et al. (2011) compared 1–3–3, 1–2–3, 1–3–2, and 1–2–2 structures using M-AMI based on the partial credit model (PCM; Masters, 1982). They found that all of the panel designs performed similarly.

In the same way, some studies have investigated the performance of variations among routing methods given a particular panel structure. For example, Kim, Chung, and Dodd (2010) compared various routing methods (e.g., M-AMI, DPI, proximity, and random) of a 1–3–3 MST based on the PCM. They found that M-AMI performed slightly better than the other methods.

Few studies, however, have investigated the performance of various panel structures with different routing methods simultaneously. For example, Zenisky (2004) compared 1–3–3, 1–2–3, 1–3–2, and 1–2–2 panel designs with DPI, proximity, and random routing procedures in order to estimate ability and determine the precision of the classification decision. A dichotomous item pool based on the three-parameter logistic IRT model was used. No substantial differences were found among these panel structures or routing methods, in that they all performed similarly in the precision of their classification decision.

To date, no study has constructed and compared various panel structures applied simultaneously via different routing methods using a polytomous item pool based on the PCM in the context of a classification testing situation, even though interaction may occur between panel structures and routing methods. Thus, in the present study we compared the performance of these panel structures with various routing methods in the classification-testing environment.

Method

For this study we used a 4 (panel structures) $\times$ 3 (routing methods) $\times$ 2 (test lengths) $\times$ 3 (passing rates) design. Thus, 72 conditions were evaluated for decision accuracy in making a binary decision (i.e., pass/fail) and test security.

Item pool

MST panels were assembled from an item bank that contained 157 items, which was obtained from a high-stakes operational testing situation. Its parameter estimates were modeled using the PCM. The PCM extends the Rasch model to include polytomous items by estimating multiple-step difficulties for each item. Each item included three, four, or five category scores. Table 1 displays the descriptive statistics for the present study's item pool. The pool included three item types and three content areas, producing nine content cells. Within the item pool, 63 % were three-category items, 18.5 % were four-category items, and 18.5 % were five-category items. The subcontent area distribution for the 157 items was 39 % area I, 37.5 % area II, and 23.5 % area III. Thus, we explored a total of nine content constraints (3 levels of category scores $\times$ 3 levels

Table 1 Descriptive statistics of the item pool

# of Step Difficulty	Parameter Estimate	N	Mean	Standard Deviation	Min	Max
2	Step Difficulty 1	99	-1.15	0.77	-2.83	0.98
	Step Difficulty 2	99	0.58	0.94	-1.81	3.57
3	Step Difficulty 1	29	-0.06	0.74	-1.07	1.50
	Step Difficulty 2	29	-0.40	0.59	-1.28	0.82
	Step Difficulty 3	29	-0.06	0.83	-1.48	1.51
4	Step Difficulty 1	29	-1.38	0.87	-3.13	0.72
	Step Difficulty 2	29	-0.60	0.72	-1.74	1.45
	Step Difficulty 3	29	-0.33	0.67	-1.30	1.33
	Step Difficulty 4	29	-0.12	0.90	-2.36	2.34

of content areas). Figure 1 provides the item pool information function for the present study.

Test lengths

Two test lengths—one test with nine items and one with 15 items—were considered for the present study. The nine-item length was selected to represent the shorter test, and the 15-item length was chosen to represent the longer test. A test length of 15 items is the maximum number of items possible to produce the 1–3–3 panel structure. In the present study, this was the largest panel structure, with at least three panels based on the capacity of the current item pool.

From the proportions of various polytomous item scores in the current item pool (three-, four-, and five-category items), converting nine polytomous items to dichotomous cases produced 23 score points. Likewise, converting 15 polytomous items produced 38 score points.

Fig. 1 Information function for the item pool

MST panel structures/construction

Four panel structures with two different test lengths were constructed automatically for the present study using a JPLEX ATA program (Park, Kim, Dodd, & Chung, 2011). This program is based on a mixed integer LP solver and applies the simplex algorithm and branch-and-bound methods (for details, see van der Linden, 2005). Figure 2 illustrates these panel designs in detail. Each panel design had three stages: the 1–3–3 panel structure had one module for the first

stage and three modules for the second and third stages (i.e., easy-, medium-, and hard-level modules). In constructing the MST panels and modules, the target TIF was centered at 0.0 (medium module) for the first stage. The target TIFs then shifted approximately by one standard deviation, centered at -1.0 (easy module) or 1.0 (hard module) on the ability continuum (i.e., the theta) for the easy and hard modules in the second and third stages. JPLEX conducted the item selection procedure to meet the requirements of the TIFs, including the location and height of the peak point and the spread of

Fig. 2 Various panel structures for a nine-item multistage test

information across the ability levels. It also took into account the proportions of content areas in the current item pool. Nine pathways were identified within each panel (i.e., a medium–medium–easy pathway, medium–medium–medium pathway, medium–medium–hard pathway, etc.). The 1–3–2 panel structure had one module for the first stage, three modules for the second stage, and two modules (i.e., easy- and hard-level modules) for the third stage. Here, six pathways were identified within each panel. The 1–2–3 panel structure had one module for the first stage, two modules for the second stage, and three modules for the third stage. Six pathways were identified within each panel. Finally, the 1–2–2 panel structure had one module for the first stage and two modules for each of the second and third stages. For this structure, four pathways were identified within each panel.

From the 157-item pool, three panels were constructed for each panel structure, with each item used only once across the panels. Three, three, and three items were assigned to each stage for the nine-item, fixed-length test condition, and five, five, and five items were assigned to each stage for the 15-item, fixed-length test condition. Additionally, each pathway was constructed to reflect the proportion of each content cell (i.e., nine content cells) of the item pool.

Routing methods

In the present study, we considered the three routing methods of M-AMI, SL-DPI, and ML-DPI. M-AMI routes examinees into the subsequent stage module that provides the highest information, given examinees' current estimated abilities. SL-DPI routes equal proportions of examinees at the current stage into subsequent stage modules. For example, in the 1–3–3 panel structure, after finishing the first stage, all of the examinees are rank-ordered according to their estimated ability. The examinees are then assigned in thirds (one-third to each condition) to the easy, medium, or hard module at the subsequent stage.

Unlike the SL-DPI routing method, ML-DPI routes the same numbers of examinees within the module at the current stage to the subsequent stage modules. For example, in the 1–3–3 panel structure, after finishing the medium module at the second stage, examinees within the medium module are rank-ordered according to their estimated abilities. About one-third of the examinees within the medium module, apiece, are then assigned to the easy, the medium, or the hard module at the third stage.

Passing rates

Three passing rates—20 %, 50 %, and 80 %—with locations for the cutoff scores based on a normal distribution of examinees were chosen for this study. These three passing rates correspond approximately to the theta points of 0.842, 0.000, and −0.842. These passing rates were selected in

order to investigate the performance of the MST designs at various points along an ability scale that represented low, middle, and high levels of ability. This was the case because tests are usually administered not only to identify examinees with higher abilities, but also to identify those with lower abilities (Thompson, 2007). Many MST studies (e.g., Hambleton & Xing, 2006; Jodoin et al., 2006; Zenisky, 2004) have used this method to choose passing rates on the basis of a normal distribution of examinees.

MST simulation

Using the 157 item parameters based on the PCM, responses were generated on the basis of the IRTGEN SAS macro (Whittaker, Fitzpatrick, Williams, & Dodd, 2003). Fifty replications with 1,000 normally simulated examinees (0, 1) were generated and used for each condition for the present study. The sample size of 1,000 is a standard sample size frequently used in many of studies that have examined polytomous MST simulations (e.g., Chen, 2010; Keng, 2008; Macken-Ruiz, 2008).

During the MST simulations, each of the simulated examinees was assigned randomly to one of three panels. After completing each stage, examinees were routed to the next-stage module according to the various routing methods on the basis of the MLE of the examinee's ability with a variable step size procedure (Koch & Dodd, 1989). According to the different conditions, fixed-length stopping rules of nine and 15 items based on the different cutoff scores (passing rates) were applied to the simulation.

Data analyses

Each simulated examinee was classified as either “passing” or “failing” by comparing the true (or known) and estimated abilities to the cutoff score points according to the three passing rates of 20 %, 50 %, and 80 %. For the present study, the correct classification was considered a true and accurate decision, which was achieved when both the true and estimated abilities were classified as either “pass” or “fail.” The false-positive rate was defined as classifications of an examinee as “passing” on the basis of estimated ability when his or her true ability should actually cause the examinee to “fail.” In the same way, a false-negative would be classifying an examinee as “failing” on the basis of estimated ability, while his or her true ability should allow the examinee to “pass.” The total classification error was then calculated as a sum of these two classification errors. Thus, all conditions were compared in terms of the following statistical values: (1) correct classification rate, (2) false-negative error rate, (3) false-positive error rate, and (4) total error rate. These values were averaged across 50 replications.

In addition, exposure control properties such as item exposure rate and pool utilization were calculated. Item

exposure rate was calculated on the basis of the number of times that a particular item was administered divided by the total number of examinees, averaged over the replications. For example, ten items having an exposure rate of .20 would indicate that, on average over the replications, these ten items were administered to 20 % of the entire group of examinees. Descriptive statistics of item exposure rates included the grand mean, the mean standard deviation, and the mean maximum value of the exposure rates.

Furthermore, pool utilization was computed on the basis of the percentage of items from among the total items constructed for each MST condition that were never administered to examinees, averaged over the replications. According to the different simulation conditions, all exposure control properties were averaged across 50 replications, with the total number of examinees being 1,000 for each replication.

Results

MST panel construction

Four panel structures were constructed according to the two chosen test length conditions. All panels in each condition were constructed similarly by obtaining the same level of TIFs. In an ideal MST design, each panel will have

Fig. 3 Examples of the actually constructed stage-level information functions for the 1–3–3 panel structure at test lengths of nine items (*upper*) and 15 items (*lower*)

approximately the same amount of information, to ensure that the test is administered fairly among the examinees. Figure 3 depicts examples of actually constructed stage-level information functions with three levels of difficulty (i.e., easy, medium, and hard) for the 1–3–3 panel structure for each test length of the present study. Figure 4 includes examples of nine actually produced pathway-level information functions of the 1–3–3 panel structure (i.e., nine pathways) for each test length. In addition, Table 2 presents the numbers of simulated examinees assigned to each pathway of the 1–3–3 panel structure according to the various routing methods using a test length of 15 items. The data show how each routing method performed differently by placing examinees into different pathways, given the 1–3–3 panel structure and test length.

MST simulations

The precision of the classification decision needs to present how accurately the tested conditions classified each

Fig. 4 Examples of the actually constructed pathway-level information functions for the 1–3–3 panel structure at test lengths of nine items (*upper*) and 15 items (*lower*). MEE = pathway with medium difficulty (M) at the first stage (S1), easy difficulty (E) at the second stage (S2), and E at the third stage (S3); MEM = pathway with M at S1, E at S2, and M at S3; MEH = pathway with M at S1, E at S2, and hard difficulty (H) at S3; MME = pathway with M at S1, M at S2, and E at S3; MMM = pathway with M at S1, M at S2, and M at S3; MMH = pathway with M at S1, M at S2, and H at S3; MHE = pathway with M at S1, H at S2, and E at S3; MHM = pathway with M at S1, H at S2, and M at S3; MHH = pathway with M at S1, H at S2, and H at S3

Table 2 Number of simulated examinees assigned to each pathway of the 1–3–3 panel structure according to different routing methods, using a test length of 15 items

		ML-DPI^c									Sum
		MEE	MEM	MEH	MME	MMM	MMH	MHE	MHM	MHH	
M-AMI ^a	MEE	78	93	51	6	2	0	0	0	0	230
	MEM	16	8	31	1	0	0	0	0	0	56
	MEH	0	0	6	1	0	1	0	0	0	8
	MME	7	4	10	50	13	0	0	0	0	84
	MMM	2	6	11	34	45	14	0	0	0	112
	MMH	0	0	3	0	14	44	0	0	0	61
	MHE	0	0	0	4	9	3	2	0	0	18
	MHM	0	0	0	4	4	8	30	0	0	46
	MHH	0	0	0	5	26	45	76	115	118	385
	Sum	103	111	112	105	113	115	108	115	118	
		ML-DPI^c									Sum
		MEE	MEM	MEH	MME	MMM	MMH	MHE	MHM	MHH	
SL-DPI ^b	MEE	103	111	63	0	0	0	0	0	0	277
	MEM	0	0	49	0	0	0	0	0	0	49
	MEH	0	0	0	0	0	0	0	0	0	0
	MME	0	0	0	49	0	0	0	0	0	49
	MMM	0	0	0	56	113	56	0	0	0	225
	MMH	0	0	0	0	0	59	0	0	0	59
	MHE	0	0	0	0	0	0	0	0	0	0
	MHM	0	0	0	0	0	0	59	0	0	59
	MHH	0	0	0	0	0	0	49	115	118	282
	Sum	103	111	112	105	113	115	108	115	118	
		SL-DPI^b									Sum
		MEE	MEM	MEH	MME	MMM	MMH	MHE	MHM	MHH	
M-AMI ^a	MEE	203	19	0	4	4	0	0	0	0	230
	MEM	45	10	0	1	0	0	0	0	0	56
	MEH	0	6	0	1	1	0	0	0	0	8
	MME	17	4	0	30	33	0	0	0	0	84
	MMM	11	8	0	11	77	5	0	0	0	112
	MMH	1	2	0	0	34	24	0	0	0	61
	MHE	0	0	0	1	15	0	0	2	0	18
	MHM	0	0	0	1	15	0	0	28	2	46
	MHH	0	0	0	0	46	30	0	29	280	385
	Sum	277	49	0	49	225	59	0	59	282	

A calculation is based on one replication. MEE = pathway with medium difficulty (M) at the first stage (S1), easy difficulty (E) at the second stage (S2), and E at the third stage (S3); MEM = pathway with M at S1, E at S2, and M at S3; MEH = pathway with M at S1, E at S2, and hard difficulty (H) at S3; MME = pathway with M at S1, M at S2, and E at S3; MMM = pathway with M at S1, M at S2, and M at S3; MMH = pathway with M at S1, M at S2, and H at S3; MHE = pathway with M at S1, H at S2, and E at S3; MHM = pathway with M at S1, H at S2, and M at S3; MHH = pathway with M at S1, H at S2, and H at S3. ^a Modified approximate maximum information. ^b Stage-level defined population intervals. ^c Module-level defined population intervals.

examinee into mutually exclusive categories (i.e., pass or fail). For the present study, the dependent measures for the precision of the classification decision included correct classification rates, false-negative error rates, false-positive error rates, and total error rates, as described.

The error rates and correct classification rates (CCRs) were calculated as percentages based on the panel structures with

routing methods and passing rates for both the nine- and 15-item test length conditions. Table 3 displays the classification accuracy for the nine-item test length conditions. For this shorter test, various panel structures given the same routing method, or all routing methods given the same panel design, produced comparable results. For example, given the M-AMI routing method, the mean CCRs for the 1–3–3 panel structure

Table 3 Comparisons of classification error and accuracy rates using the test length of nine items

Panel Design	Routing Method	Passing Rate (%)	CCR (%)	FNER (%)	FPER (%)	TER (%)
1–3–3	M-AMI ^a	20	90.86	4.03	5.11	9.14
		50	88.83	6.23	4.94	11.17
		80	92.15	4.39	3.46	7.85
	SL-DPI ^b	20	90.84	4.04	5.12	9.16
		50	88.64	6.18	5.18	11.36
		80	92.14	4.15	3.71	7.86
	ML-DPI ^c	20	90.58	3.75	5.67	9.42
		50	88.33	6.31	5.36	11.67
		80	91.61	4.91	3.48	8.39
1–2–3	M-AMI ^a	20	90.95	4.03	5.02	9.05
		50	88.77	5.97	5.26	11.23
		80	92.07	4.05	3.88	7.93
	SL-DPI ^b	20	91.04	4.17	4.79	8.96
		50	88.64	5.89	5.47	11.36
		80	92.12	4.06	3.82	7.88
	ML-DPI ^c	20	90.77	3.70	5.53	9.23
		50	87.96	6.31	5.73	12.04
		80	91.62	4.57	3.81	8.38
1–3–2	M-AMI ^a	20	90.94	4.10	4.96	9.06
		50	88.08	6.34	5.58	11.92
		80	92.15	4.49	3.36	7.85
	SL-DPI ^b	20	90.88	3.97	5.15	9.12
		50	87.99	6.32	5.69	12.01
		80	92.16	4.28	3.56	7.84
	ML-DPI ^c	20	90.74	4.31	4.95	9.26
		50	88.06	6.24	5.70	11.94
		80	91.43	4.99	3.58	8.57
1–2–2	M-AMI ^a	20	91.09	4.13	4.78	8.91
		50	87.92	5.79	6.29	12.08
		80	92.14	4.13	3.73	7.86
	SL-DPI ^b	20	91.09	4.15	4.76	8.91
		50	87.74	6.34	5.92	12.26
		80	92.15	4.15	3.70	7.85
	ML-DPI ^c	20	90.90	4.56	4.54	9.10
		50	87.95	6.32	5.73	12.05
		80	91.82	4.28	3.90	8.18

All statistics were averaged across 50 replications. Each replication contained 1,000 observations. CCR = correct classification rate; FNER = false-negative error rate; FPER = false-positive error rate; TER = total error rate. ^a Modified approximate maximum information. ^b Stage-level defined population intervals. ^c Module-level defined population intervals.

ranged from 88.83 % to 92.15 %; the mean CCRs for 1–2–3 ranged from 88.77 % to 92.07 %; the mean CCRs for 1–3–2 ranged from 88.08 % to 92.15 %; and the mean CCRs for the 1–2–2 panel design ranged from 87.92 % to 92.14 % across the three passing-rate conditions. In a like manner, in terms of the 1–3–3 panel design, the mean CCRs for the M-AMI

ranged from 88.83 % to 92.15 %; the mean CCRs for the SL-DPI ranged from 88.64 % to 92.14 %, and the mean CCRs for the ML-DPI ranged from 88.33 % to 91.61 % across the three passing-rate conditions.

Table 4 Comparisons of classification error and accuracy rates using the test length of 15 items

Panel Design	Routing Method	Passing Rate (%)	CCR (%)	FNER (%)	FPER (%)	TER (%)
1–3–3	M-AMI ^a	20	92.86	3.05	4.09	7.14
		50	90.54	5.09	4.37	9.46
		80	93.25	3.69	3.06	6.75
	SL-DPI ^b	20	92.72	3.12	4.16	7.28
		50	90.40	5.35	4.25	9.60
		80	93.36	3.62	3.02	6.64
	ML-DPI ^c	20	92.51	3.39	4.10	7.49
		50	90.51	5.23	4.26	9.49
		80	93.12	3.96	2.92	6.88
1–2–3	M-AMI ^a	20	92.88	2.99	4.13	7.12
		50	90.38	5.13	4.49	9.62
		80	93.29	3.67	3.04	6.71
	SL-DPI ^b	20	92.84	3.04	4.12	7.16
		50	90.22	5.07	4.71	9.78
		80	93.40	3.55	3.05	6.60
	ML-DPI ^c	20	92.65	3.27	4.08	7.35
		50	90.17	5.00	4.83	9.83
		80	93.28	3.87	2.85	2.72
1–3–2	M-AMI ^a	20	92.88	3.03	4.09	7.12
		50	90.43	5.25	4.32	9.57
		80	93.31	3.67	3.02	6.69
	SL-DPI ^b	20	92.75	3.08	4.17	7.25
		50	90.44	5.06	4.50	9.56
		80	93.36	3.62	3.02	6.64
	ML-DPI ^c	20	92.49	3.43	4.08	7.51
		50	90.55	4.98	4.47	9.45
		80	93.12	3.87	3.01	6.88
1–2–2	M-AMI ^a	20	92.88	2.98	4.14	7.12
		50	90.09	5.38	4.53	9.91
		80	93.40	3.51	3.09	6.60
	SL-DPI ^b	20	92.87	3.00	4.13	7.13
		50	89.97	5.36	4.67	10.03
		80	93.41	3.54	3.05	6.59
	ML-DPI ^c	20	92.78	3.30	3.92	7.22
		50	90.20	4.96	4.84	9.80
		80	93.25	3.58	3.17	6.75

All statistics were averaged across 50 replications. Each replication contained 1,000 observations. CCR = correct classification rate; FNER = false-negative error rate; FPER = false-positive error rate; TER = total error rate. ^a Modified approximate maximum information. ^b Stage-level defined population intervals. ^c Module-level defined population intervals.

As with the nine-item test conditions, all panel structures or routing methods produced comparable results for the 15-item test length condition (see Table 4). For example, in terms of the M-AMI routing method, the 1–3–3 structure produced mean CCRs ranging from 90.54 % to 93.25 %; the 1–2–3 structure produced CCRs ranging from 90.38 % to 93.29 %; the 1–3–2 structure produced CCRs ranging from 90.43 % to 93.31 %; and the 1–2–2 structure produced CCRs ranging from 90.09 % to 93.40 % across the three passing rate conditions. In a like manner, given the 1–3–3 panel structure, the mean CCRs for the M-AMI routing method ranged from 90.54 % to 93.25 %; the mean CCRs for the SL-DPI routing method ranged from 90.40 % to 93.36 %; and the mean CCRs for the ML-DPI routing method ranged from 90.51 % to 93.12 %, across different passing rate conditions.

For all of the 15-item test length panel designs and routing methods, as with the nine-item test length, the passing rate of 50 % produced the lowest mean CCRs (see Table 4). This is the case because the 50 % passing rate places the cutoff scores on the ability scale at a point where most of the examinees cluster in a normal distribution (e.g., Jodoin, 2003; Zenisky, 2004). Generally, the 15-item test length produced higher mean CCRs than did the nine-item test length by approximately 1.10 % to 2.49 %, given the panel design, routing method, and passing rate conditions.

Pool utilization, frequency distribution, and descriptive statistics of item exposure rates were calculated as exposure control properties for the two different test lengths. For example, Table 5 displays the exposure control properties when the passing rate was set at 20 %. It contains how many items fell into the mean exposure rate categories. It also presents mean standard deviations, mean maximum, and grand means of exposure rates and pool utilization indices averaged across 50 replications. Computations of the exposure control properties of the MST conditions were based only on the proportions of the entire item pool used to construct the MST panels.

For the fixed-length tests, the mean exposure rate was calculated on the basis of the ratio of test length to pool size when the test length was fixed (Chen, Ankenmann, & Spray, 2003). Thus, the grand means of item exposure rates varied across conditions, depending on the size of the item pool used to construct the panels. The grand means of item exposure rates varied from .14 (e.g., 9/63 or 15/105) to .20 (e.g., 9/45 or 15/75), according to the different panel structures. Moreover, because each item was used only once across the three panels, each of which consisted of five (i.e., 1–2–2) to seven (i.e., 1–3–3) modules, and each panel was assigned randomly to each examinee, the mean maximum exposure rate was .337 across all conditions. In other words, certain items were exposed or administered to about 34 % of the entire group of examinees. The pool utilization

rates were excellent throughout all conditions, because all of the items used to construct the MST were administered to examinees.

Discussion

Many design components can be considered for constructing and administering MST, such as test information, panel structures, passing rates, ability estimations, and routing methods. The goal of the present research was to construct and investigate various MST panel structures interacting with routing methods from an item pool based on PCM. Three routing methods were used for the present study. Among these, the two DPI methods were relatively easy to implement, but these methods did not consider the optimal matching of examinees' current ability to the subsequent module information. The M-AMI method used the currently estimated ability and information of the modules, but it might result in overexposing certain modules (Zenisky, 2004).

Generally, with the same test length condition, all routing methods, given the same panel structure, worked similarly. In addition, all panel structures, given the same routing method with the same test length condition, performed equally well in terms of the precision of the classification decision across all conditions. Selecting either two or three modules in the second or third stages of testing seemed to make no practical difference to the results. The reason that the routing methods or panel structures did not make a significant difference for the present study might be that the constructed TIFs of modules at the given stage (i.e., easy–hard or easy–medium–hard) of the various MST panel structures were not separated sharply. In other words, they were constructed to somewhat overlap one another, so as to cover the entire ability range (i.e., theta of -4.0 to 4.0) established in the present study.

For the present study, the differences in mean CCRs among the conditions of the routing methods and panel structures were less than 2 %, given the test length and passing rate conditions. According to Jodoin (2003), testing programs with low to moderate consequences may not have practical impacts that dictate which design to use within the 2 % of the practical difference range. In a high-stakes testing situation, however, such as selecting medical doctors, test administrators need to consider which design is most suitable for their programs, to ensure the most accurate results.

As expected, when the test length increased, all of the panel designs, given the routing method condition (or vice versa), produced better results in classification accuracy than did the shorter test length conditions. These results correspond to a well-known characteristic of the test:

Table 5 Pool utilization and exposure rates for MST

	Nine Items									15 Items								
	1-3-3			1-3-2			1-2-3			1-2-2			1-3-3			1-3-2		
	MA	SD	MD	MA	SD	MD	MA	SD	MD	MA	SD	MD	MA	SD	MD	MA	SD	MD
PS ^a	.63	.63	.63	.54	.54	.54	.54	.54	.54	.45	.45	.45	.105	.105	.105	.15	.15	.15
ER	.9	.9	.9	.9	.9	.9	.9	.9	.9	.9	.9	.9	.15	.15	.15	.15	.15	.15
.31-.40	.9	.9	.9	.9	.9	.9	.9	.9	.9	.9	.9	.9	.15	.15	.15	.15	.15	.15
.21-.30	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
.11-.20	.33	.54	.54	.33	.45	.45	.34	.45	.45	.33	.36	.36	.53	.90	.90	.52	.75	.75
.06-.10	.17	.0	.0	.8	.0	.0	.8	.0	.0	.0	.0	.0	.26	.0	.0	.0	.0	.0
.01-.05	.4	.0	.0	.4	.0	.0	.0	.0	.0	.0	.0	.0	.11	.0	.0	.1	.0	.0
NA	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
ERA	.14	.14	.14	.17	.17	.17	.17	.17	.17	.20	.20	.20	.14	.14	.14	.17	.17	.17
ERSD	.087	.078	.078	.089	.079	.079	.083	.079	.079	.071	.067	.067	.090	.078	.078	.085	.079	.079
ERM	.337	.337	.337	.337	.337	.337	.337	.337	.337	.337	.337	.337	.337	.337	.337	.337	.337	.337
% of pool NA	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0

This table only includes conditions related to the passing rate of 20 %. All statistics were averaged across 50 replications, and each replication contained 1,000 observations. MA, modified approximate maximum information; SD, stage-level defined population intervals; MD, module-level defined population intervals; PS, pool size; ER, exposure rate; NA, not administered; ERA, exposure rate average; ERSD, exposure rate standard deviation; ERM, exposure rate maximum. ^a MST used only the proportion of the entire item pool in panel assembly for each test length condition.

Longer test lengths lead to an increase in decision consistency or accuracy (Crocker & Algina, 1986).

In terms of the exposure control properties, because each item was used only one time across three panels, and one of the panels was assigned randomly to the simulated examinee according to different panel structures, the mean maximum exposure rate was less than .35 across all conditions. Unlike the M-AMI method, the two DPI methods performed in a similar way, because these methods routed examinees to the subsequent-stage module according to predefined proportions. The M-AMI method yielded slightly higher mean exposure rate standard deviations than did the two DPI methods. Finally, all of the MST conditions achieved excellent pool utilization rates by using all of the items available to construct the panels.

In the present study, we investigated multiple panel designs simultaneously, applied with various routing methods for MST based on the PCM. The present results confirmed that various MST panel designs with different routing methods performed well using the item pool based on the PCM. Different routing methods, therefore, can be applied flexibly to the specified panel structure. Furthermore, smaller panel structures, such as the 1–2–2 panel design, with various routing methods produced similar results relative to other panel structures, even though the two-module structure in the second- and the third-stage MST design has only one possible adaption point based on estimated abilities. This implies that MST panels could be constructed more economically—that is, using fewer items. These results might be useful for testing programs interested in using MST for classification decisions with a smaller item bank.

To generalize the present study's results, future research should consider various other conditions. For example, future research might use a larger item pool with more items than in the present study, such that more panels could be constructed for each condition. This would reduce the exposure rate even further (e.g., the mean maximum exposure rates) than has been found in the present study. In doing so, such extended studies would also be able to investigate a larger variety of test lengths and panel designs. In the present study, the constructed TIFs for modules within each level of difficulty were spread widely across the ability scale. Thus, a future study might try to minimize overlaps across ability scales by constructing TIFs from MST modules that are more distinctive, in order to route examinees to the proper modules on the basis of their ability estimates. This would be especially useful for the M-AMI method, which is based on TIFs. For the present study, all of the panel structures featured three stages. Thus, it will be interesting to see how different routing methods would perform differently using two-stage panel structures

(e.g., 1–3 or 1–4). Finally, we did not reflect the examinees' actual scenario with MST, particularly their ability to review the items and correct their answers within a module before moving to the next stage. Future research could incorporate this MST administrative feature into the study design.

References

- Betz, N. E., & Weiss, D. J. (1974). *Simulation studies of two-stage ability testing (Research Report No. 74-4)*. Minneapolis, MN: University of Minnesota, Department of Psychology, Psychometric Methods Program.
- Breithaupt, K., & Hare, D. R. (2007). Automated simultaneous assembly of multistage testlets for a high-stakes licensing examination. *Educational and Psychological Measurement*, 67, 5–20. doi:10.1177/0013164406288162
- Chen, L. Y. (2010). *An investigation of the optimal test design for multi-stage test using the generalized partial credit model*. Unpublished doctoral dissertation, University of Texas, Austin.
- Chen, S., Ankenmann, R. D., & Spray, J. A. (2003). The relationship between item exposure and test overlap in computerized adaptive testing. *Journal of Educational Measurement*, 40, 129–145. doi:10.1111/j.1745-3984.2003.tb01100.x
- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York, NY: CBS College Publishing.
- Davis, L. L., & Dodd, B. G. (2003). Item exposure constraints for testlets in the verbal reasoning section of the MCAT. *Applied Psychological Measurement*, 27, 335–356. doi:10.1177/0146621603256804
- 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. doi:10.1207/s15324818ame1903_4
- Hendrickson, A. (2007). An NCME instructional module on multistage testing. *Educational Measurement: Issue and Practice*, 26, 44–52. doi:10.1111/j.1745-3992.2007.00093.x
- Jodoin, M. G. (2003). *Psychometric properties of several computer-based test designs with ideal and constrained item pools*. Unpublished doctoral dissertation, University of Massachusetts, Amherst.
- Jodoin, M. G., Zenisky, A., & 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. doi:10.1207/s15324818ame1903_3
- Keng, L. (2008). *A comparison of the performance of testlet-based computer adaptive tests and multistage tests*. Unpublished doctoral dissertation, University of Texas, Austin.
- Kim, J., Chung, H., & Dodd, B. G. (2010, April). *Comparing routing methods in the multistage test based on the partial credit model*. Paper presented at the annual meeting of the American Educational Research Association, Denver, CO.
- Kim, J., Chung, H., Park, R., & Dodd, B. G. (2011, April). *A comparison of panel designs in the multistage test based on the partial credit model*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.
- Kim, H., & Plake, B. S. (1993, April). *Monte Carlo simulation comparison of two-stage testing and computerized adaptive testing*. Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta, GA.
- Koch, W. R., & Dodd, B. G. (1989). An investigation of procedures for computerized adaptive testing using partial credit scoring.

- Applied Measurement in Education*, 2, 335–357. doi:10.1177/014662160528007
- Linn, R. L., Rock, D. A., & Cleary, T. A. (1969). The development and evaluation of several programmed testing methods. *Educational and Psychological Measurement*, 29, 129–146. doi:10.1177/001316446902900109
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.
- 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., Brumfield, T., & Breithaupt, K. (2006). A testlet assembly design for adaptive multistage tests. *Applied Measurement in Education*, 19, 189–202. doi:10.1207/s15324818ame1903_2
- Luecht, R. M., & Nungester, R. J. (1998). Some practical examples of computer-adaptive sequential testing. *Journal of Educational Measurement*, 35, 229–249. doi:10.1111/j.1745-3984.1998.tb00537
- Macken-Ruiz, C. L. (2008). *A comparison of multi-stage and computerized adaptive tests based on the generalized partial credit model*. Unpublished doctoral dissertation, University of Texas, Austin.
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149–174.
- Park, R., Kim, J., Dodd, B. G., & Chung, H. (2011). JPLeX: Java simplex implementation with branch-and-bound search for automated test assembly. *Applied Psychological Measurement*, 35, 643–644. doi:10.1177/0146621610392912
- Patsula, L. (1999). *A comparison of computerized adaptive testing and multi-stage testing*. Unpublished doctoral dissertation, University of Massachusetts, Amherst.
- 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, Canada.
- Thompson, N. A. (2007). *A comparison of two methods of polytomous computerized classification testing for multiple cutscores*. Unpublished doctoral dissertation, University of Minnesota, Twin Cities.
- van der Linden, W. J. (2005). *Linear models for optimal test design*. New York, NY: Springer.
- Wainer, H. (1995). Precision and differential item functioning on a testlet-based test: The 1991 Law School Admissions Test as an example. *Applied Measurement in Education*, 8, 157–187. doi:10.1207/s15324818ame0802_4
- Whittaker, T. A., Fitzpatrick, S. J., Williams, N. J., & Dodd, B. G. (2003). IRTGEN: A SAS macro program to generate known trait scores and item response for commonly used item response theory models. *Applied Psychological Measurement*, 27, 299–300. doi:10.1177/0146621603027004005
- Xing, D. (2001). *Impact of several computer-based testing variables on the psychometric properties of credentialing examinations*. Unpublished doctoral dissertation, University of Massachusetts, Amherst.
- 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.