Journal of Computerized Adaptive Testing

Volume 2 Number 2

February 2014

A Comparison of Multi-Stage and Linear Test Designs for Medium-Size Licensure and Certification Examinations

Bradley G. Brossman and Robin A. Guille

DOI 10.7333/1402-0202018

The *Journal of Computerized Adaptive Testing* is published by the International Association for Computerized Adaptive Testing
www.iacat.org/jcat

ISSN: 2165-6592

©2014 by the Authors. All rights reserved.
This publication may be reproduced with no cost for academic or research use.
All other reproduction requires permission from the authors;
if the author cannot be contacted, permission can be requested from IACAT.

Editor
David J. Weiss, *University of Minnesota, U.S.A*

Associate Editor
G. Gage Kingsbury
Psychometric Consultant, U.S.A.

Associate Editor
Bernard P. Veldkamp
University of Twente, The Netherlands

Consulting Editors

John Barnard
EPEC, Australia
Juan Ramón Barrada
Universidad de Zaragoza, Spain
Kirk A. Becker
Pearson VUE, U.S.A.
Barbara G. Dodd
University of Texas at Austin, U.S.A.
Theo Eggen
Cito and University of Twente, The Netherlands
Andreas Frey
Friedrich Schiller University Jena, Germany
Kyung T. Han
Graduate Management Admission Council, U.S.A.
Wim J. van der Linden
CTB/McGraw-Hill, U.S.A.
Alan D. Mead
Illinois Institute of Technology, U.S.A.
Mark D. Reckase
Michigan State University, U.S.A.
Barth Riley
University of Illinois at Chicago, U.S.A.
Otto B. Walter
University of Bielefeld, Germany
Wen-Chung Wang
The Hong Kong Institute of Education
Steven L. Wise
Northwest Evaluation Association, U.S.A.

Technical Editor
Martha A. Hernández

International Association for Computerized Adaptive Testing
IACAT
Advancing the Science and Practice of Human Assessment

Journal of Computerized Adaptive Testing

Volume 2, Number 2, February 2014
DOI 10.7333/1402-020218
ISSN 2165-6592

A Comparison of Multi-Stage and Linear Test Designs for Medium-Size Licensure and Certification Examinations

Bradley G. Brossman and Robin A. Guille
American Board of Internal Medicine

Many large-scale testing organizations currently use multi-stage testing (MST) for their examinations. The MST design implements the test in several stages, where one module is administered per stage. Successive stages in MST might vary in difficulty depending on the estimated ability level of the examinee. Although several studies have been conducted to compare the performance of MST to the traditional linear test design, all of the investigations known to date have incorporated simulation studies that capitalize on large sample size requirements in order to reduce the number of replicated datasets. As a result, although the statistics under investigation have been estimated with reasonable stability, these studies have been better suited to investigate the performance of MST for large-scale examinations as opposed to small- or medium-size examinations. The purpose of this research was to conduct a series of studies based on simulated datasets for medium-size medical certification examinations. The results confirmed that more accurate ability estimates and more accurate and consistent pass-fail decisions are obtained under the MST design compared to the traditional linear design for these examinations.

Keywords: *adaptive testing, multi-stage testing, licensure and certification, testing in the professions, testlets.*

Adaptive testing, which tailors a test to match the characteristics of the examinee, has become a common form of test administration during the past few decades (Armstrong & Little, 2003; Guille et al., 2011; Hambleton & Xing, 2006; Jodoin, Zenisky, & Hambleton, 2006; Luecht, Brumfield, & Breithaupt, 2006; Luecht & Sireci, 2011; Xing & Hambleton, 2004; Zenisky, 2004). Adaptive tests have been demonstrated to be more efficient than traditional linear, or "fixed-form," tests in that fewer items are required to obtain the same amount of measurement precision. Similarly, adaptive tests that contain the same number of items as traditional linear tests typically result in more precise estimates of examinee ability (Drasgow, Luecht, & Bennett, 2006).

A variety of adaptive test models have been developed to meet the varied and diverse needs of different testing organizations (see Luecht and Sireci, 2011 for a summary of the various

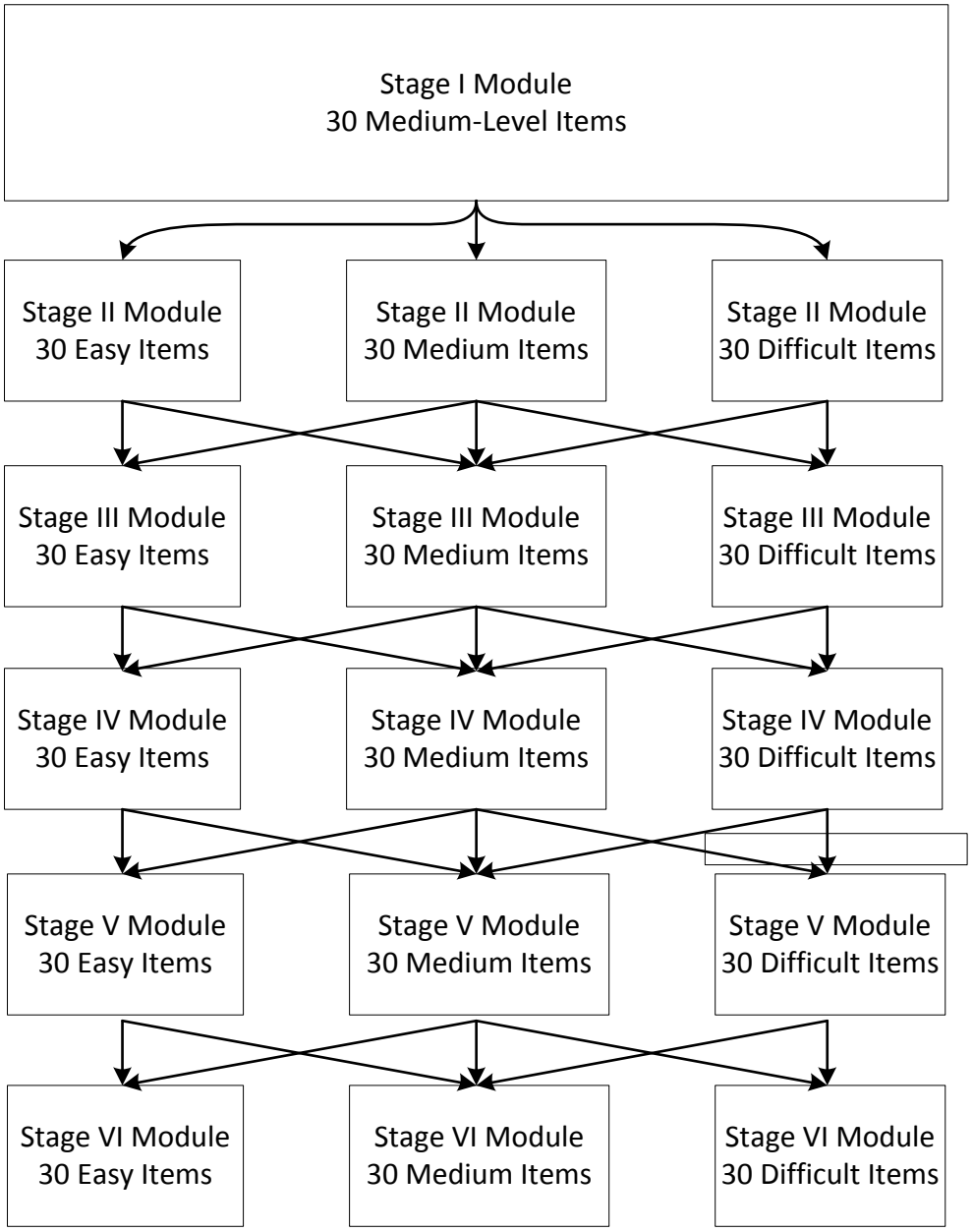
adaptive models that have been developed). These models primarily differ in regard to the level at which the adaptation occurs (Zenisky, 2004). For some models such as computerized adaptive testing (CAT), individual items are selected one at a time based on examinee performance to all previous items. For other models, such as multi-stage testing (MST), sets of items are selected based on examine performance to previous sets of items. The latter family of models more closely resembles traditional linear tests, in that all test forms that will ultimately be administered are known prior to test administration (Armstrong & Little, 2003; Guille et al., 2011; Luecht & Sireci, 2011; Zenisky, 2004).

MST has been commonly used in practice due to its capability for quality control. Similar to a linear design in that every test panel that will ultimately be administered is specified in advance, MST methods allow for each test panel to be reviewed for statistical specifications, content specifications, and other features such as graphics and audio-visual components prior to test administration (Armstrong & Little, 2003; Guille et al., 2011; Luecht et al., 2006; Luecht & Sireci, 2011; Zenisky, 2004). This allows for higher quality control monitoring than the "on-the-fly" methods employed in CAT, yet with the added advantage of an adaptive component compared to traditional linear methods.

To conduct MST, items are first grouped into clusters, or *modules*, prior to administration such that each module satisfies given statistical and content specifications. These modules form the building blocks for the MST design. Typically, each module is constructed such that the module-level content specifications match the content specifications for the overall test. The statistical specifications, however, are intended to vary across modules such that different modules of items are associated with different levels of difficulty. For example, testing organizations that use MST typically create *difficult* modules (i.e., modules that are intended for examinees who perform very well on the test), *easy* modules (i.e., modules that are intended for examinees who do not perform well on the test), and *medium* modules (i.e., modules that are intended for examinees who perform neither exceedingly well nor exceedingly poorly on the test). After each module is created, test developers review the module to ensure that the content and statistical specifications have been met.

During a typical MST administration, all examinees are first administered the same module of items, which is usually specified to be near the medium ability level in terms of item difficulty. Based on the responses to this first module of items, an interim ability score is estimated for each examinee. Then, based on pre-specified decision rules, examinees are administered a Stage 2 module comprised of either difficult, medium, or easy items in accordance with their interim Stage 1 ability estimate. If the test is comprised of only two stages, the test is finished after Stage 2. However, tests in the MST framework are often structured to have more than two stages (Zenisky, 2004). In this situation, after the Stage 2 modules are administered, ability scores are re-estimated for each examinee based on all responses to the Stage 1 and Stage 2 items and examinees are once again routed into Stage 3 difficult, easy, or medium modules based on their re-estimated interim ability estimates. This process continues until all stages have been administered. After the test is finished, ability scores for each examinee are estimated based on responses to all administered items. Figure 1 shows the pathways for a MST design with six stages, with each stage comprised of 30 items.

Figure 1. Example MST Diagram with Routing Rules



Prior Research on MST

A number of studies have been conducted to compare both the qualitative and quantitative performance of MST with other test designs (Armstrong & Little, 2003; Guille et al., 2011; Hambleton & Xing, 2006; Jodoin, Zenisky, & Hambleton, 2002; Jodoin et al., 2006; Luecht et al., 2006; Luecht & Burgin, 2003; Luecht & Sireci, 2011). Most of these investigations are based on tests administered by testing organizations in the licensure and certification industry (Guille et al., 2011; Hambleton & Xing, 2006; Jodoin et al., 2002, 2006; Luecht et al., 2006; Luecht & Sireci, 2011). For example, Jodoin et al. (2002, 2006) compared a linear test design with various MST designs by using operational data used to make pass-fail decisions from a credentialing

agency. The authors concluded that the performance of the MST and linear designs were comparable.

Hambleton and Xing (2006) investigated the effect of targeting optimal and non-optimal test information functions (TIFs) for linear, CAT, and MST test designs. The operational data used in the study came from a credentialing test used for making pass-fail decisions. In this study, optimization was determined by cut score (i.e., targeting TIFs based on the cut score) and by ability distribution (i.e., targeting TIFs based on the mean of the ability distribution). The authors concluded that although the MST design performed only slightly better than the linear design in regard to psychometric criteria, the MST design might be the preferred design based on qualitative advantages such as better item bank utilization, candidate preference, and diagnostic feedback (Hambleton & Xing, 2006).

In a more recent study, Guille et al. (2011) investigated asymmetric termination (i.e., allowing high performing examinees to terminate the test early) and subscore reliability under the MST and linear designs for a medical certification test. The authors concluded that although favorable asymmetric termination results were obtained under the MST design, the results of the subscore standard error estimates were mixed.

Overall, a number of studies have been conducted to evaluate different aspects of MST and linear test designs under a variety of conditions, mostly within the sphere of licensure and certification tests. Whereas these investigations have historically used simulation studies to evaluate performance, nearly all of the investigations known to date have included very few simulated replications by which performance could be evaluated. To compensate for the number of replications, these studies typically incorporate larger sample sizes per replication. In reference to the number of replications included in their study, Hambleton and Xing (2006) note:

> This sample size was large enough to produce very stable estimates of statistics of interest. Preliminary research suggested that the statistics of interest in this study would vary by less than 0.002 when the sample sizes were as large as 5,000, and thus replication was not necessary (p. 225).

Although it is possible to evaluate the performance of the MST design with very few simulated replications when the sample size is large, it might be of interest to broaden the generalizability of these studies by including fewer examinees per replication. In this case (i.e., fewer examinees per replication), the simulation design is more consistent with small- and medium-sized tests often found in practice. Naturally, more replications would be required to attain the same precision as previous studies when smaller sample sizes per replication are used.

The purpose of the present study was to expand the prior research base concerning the performance of MSTs by conducting simulation studies that incorporated more replications and fewer examinees per replication. From this perspective, the present study was an attempt to provide comparative information between the MST and linear designs for medical certification tests with small- and medium-sized samples.

Method

A series of simulation studies was conducted to compare the performance of MST and linear testing in regard to estimation accuracy (i.e., how accurately ability was estimated), decision accuracy, and decision consistency based on data from a medical certification test. The results were also used to observe and describe early termination patterns for examinees who clearly passed the simulated tests. The studies were conducted under a variety of conditions in order for the results to be as generalizable as possible. The simulation conditions were selected so as to approx-

imate the certification test results as closely as possible, while at the same time maintaining breadth so as to generalize the results to as wide a population as possible.

Test Assembly

Each simulation study incorporated three-parameter logistic (3PL) item response theory (IRT) parameters from the certification test item bank. First, target TIFs were calculated to specify the statistical targets for the easy, medium, and difficult modules. These targets were determined by grouping the certification test ability (θ) estimates into three groups: the lower quartile, the interquartile, and the upper quartile. For each group separately, the average amount of information that each item contributed was calculated (i.e., item information was calculated at each θ estimate, and these item information values were then averaged across the θ estimates within each of the three groups). Items were then rank ordered according to the average amount of information provided across the group. For example, for the lower quartile group, the items were rank ordered according to the average amount of information provided for this group. Similar rank orderings were calculated for the interquartile group and for the upper quartile group. Naturally, easier items provided greater information for the lower quartile group and were near the top of the rank-ordering for that group, whereas difficult items provided greater information for the upper quartile group and were near the top of the rank-ordering for that group.

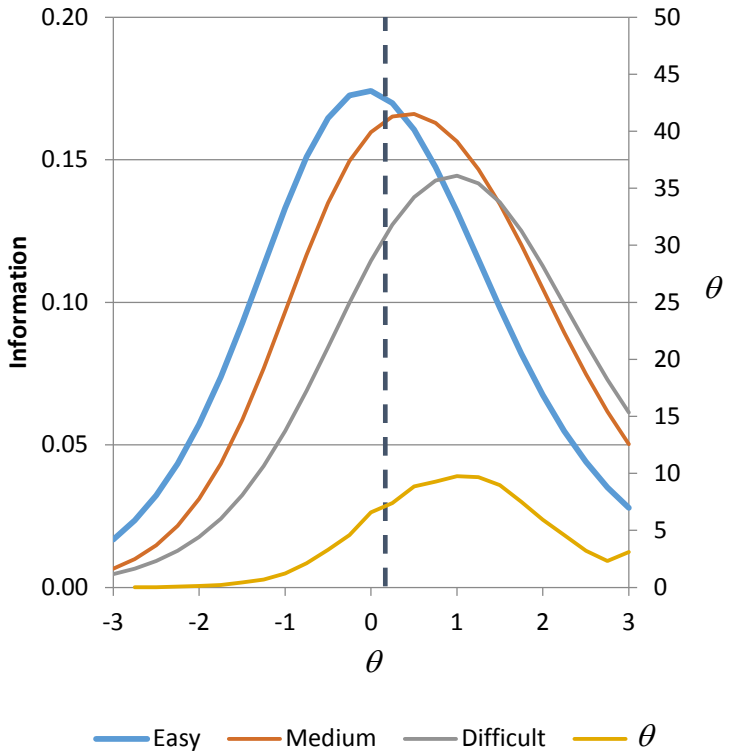
After the rank-ordering, the top one-third of highest information items were selected for each group, and target TIFs were calculated (corresponding to the easy modules, medium modules, and difficult modules, respectively) based on this top one-third of most informative items. Although arbitrary, the top one-third of items was selected so as to yield TIFs that discriminated among the θ groups. For example, if all items were selected for each group (as opposed to the top one-third of items), the target TIFs would be identical for the easy, medium, and difficult modules. As the percentage of most informative items decreased, the target TIFs became more distinct. At the same time, the top one-third of items were selected so as not to be too discriminating, which would result in creating target TIFs that might be difficult to match in subsequent test administrations. For example, depending on the size of the item bank, if the items contributing to target TIFs were too selective, it would be difficult to maintain those information levels across subsequent administrations. Given that many testing organizations control for statistical parallelism across forms by matching the same target TIFs across subsequent administrations, it seemed prudent to use the top one-third of most informative items when creating target TIFs; the one-third criterion allowed for a differentiation in target TIFs across easy, medium, and difficult groups, while at the same time not being too selective so as to allow for reproducibility across subsequent administrations.

In summary, the target TIFs were calculated by maximizing the test information for each of the three groups (representative of the easy, medium, and difficult module target TIFs). The final target TIFs, along with the ability distribution for the certification test, are shown in Figure 2 (note that the TIFs are calculated as the average of the item information as opposed to the sum of the item information). The dotted line represents the cut score value of 0.165. Table 1 contains the summary statistics for the 3PL IRT parameters corresponding to each module. It can be seen from Figure 2 that the easy TIF has slightly more information than the medium TIF, which in turn has slightly more information than the difficult TIF. This is a byproduct of the item bank containing more easy than difficult items. The actual TIFs used in this study are not shown, as they deviated only slightly from the target TIFs.

Once target TIFs were determined for each module (easy, medium, and difficult), 480 items were selected from the entire bank using automated test assembly (ATA) procedures to create the

simulated tests. The ATA procedure used the Mixed Integer Programming algorithm using IBM ILOG CPLEX software and took into account the target TIFs, content constraints, and item exposure, along with other criteria, to create each module. For this particular study, content constraints were set on each of five content domains for each module as follows: Cardiology (9 items, 30%); Pulmonary Disease (6 items, 20%); Gastroenterology (5 items, 17%); Infectious Disease (5 items, 17%); and Endocrinology (5 items, 17%). 30 items were selected to appear on the Stage 1 medium module, which all examinees complete. 30 items were also selected for each of the Stages 2, 3, 4, 5, and 6 easy, medium, and difficult modules, respectively. The same panel of items was used for every replication in this study.

Figure 2. Target TIFs and θ Distribution



Several issues arise when the MST design is implemented for small- and medium-size tests, especially when the tests are used to make pass-fail decisions in the licensure and certification fields. Although some of these issues affect small- or medium-size tests in general, regardless of whether an adaptive or linear design was used (e.g., item parameter estimation, standard setting, controlling for item parameter drift), other issues surrounding small samples are specific to the MST design.

For example, one issue that arises is how to maintain parallel TIFs for each pathway across subsequent test administrations. Although this might also be a concern when using a linear test design, the number of target TIFs increases as the number of possible pathways increases under the MST design. If the target TIFs are set too high or too low, or if the target TIFs are set to be too easy or too difficult, these TIFs might be difficult to maintain across several administrations when using a small item bank. In this study, target TIFs were calculated based on the top

Table 1. Descriptive Statistics of IRT Parameters by Module and Stage

Parameter and Stage	Easy		Medium		Difficult	
	Mean	*SD*	Mean	*SD*	Mean	*SD*
a Parameter						
Stage 1	–	–	0.63	0.16	–	–
Stage 2	0.63	0.16	0.62	0.15	0.55	0.14
Stage 3	0.62	0.12	0.62	0.11	0.56	0.15
Stage 4	0.62	0.16	0.62	0.13	0.55	0.11
Stage 5	0.62	0.16	0.63	0.14	0.55	0.14
Stage 6	0.62	0.11	0.63	0.12	0.56	0.15
b Parameter						
Stage 1	–	–	0.33	0.78	–	–
Stage 2	-0.40	0.76	0.24	0.61	0.68	0.55
Stage 3	-0.38	0.55	0.27	0.66	0.72	0.62
Stage 4	-0.37	0.56	0.30	0.71	0.67	0.54
Stage 5	-0.36	0.68	0.37	0.77	0.71	0.58
Stage 6	-0.38	0.49	0.34	0.68	0.71	0.56
c Parameter						
Stage 1	–	–	0.20	0.01	–	–
Stage 2	0.20	0.00	0.20	0.01	0.20	0.01
Stage 3	0.20	0.01	0.20	0.01	0.20	0.01
Stage 4	0.20	0.01	0.20	0.02	0.20	0.01
Stage 5	0.20	0.01	0.20	0.01	0.20	0.01
Stage 6	0.20	0.00	0.20	0.01	0.20	0.00

one-third of highest information items for each of three ability groups. This “one-third criterion” yielded reproducible target TIFs (for use across several administrations), given the size of the particular item bank used in this study.

Along similar lines, the length of the test has obvious effects on creating parallel target TIFs across subsequent administrations. Most licensure and certification tests are quite long due to reliability and validity considerations for such high-stakes assessments. For example, the current certification test under investigation is comprised of 180 items. If the test assembly procedure takes into account item exposure constraints—along with content and statistical constraints—the length of the test certainly has an impact on producing parallel pathways across adjacent administrations. Longer tests have a greater number of items that are administered, which increases item exposure and reduces the number of possible items to be selected for each pathway in future administrations.

Lastly, small- and medium-size tests—especially in the licensure and certification industry—are often created from item banks where the distribution of the item difficulties and the ability distribution are not perfectly aligned. For example, in this study Figure 2 shows that the item

bank tended to be slightly easy compared to the ability distribution, which is why the easy target TIF was slightly higher than the medium target TIF, and why the medium target TIF was slightly higher than the difficult target TIF. The fact that the item bank tended to be slightly easy compared to the ability distribution is not a flaw in the item bank (although the information for the easy, medium, and difficult target TIFs will vary as a result). Rather, the items in this bank were specifically designed to be near the cut score in terms of difficulty level. Given the high pass rates for this particular certification test (and for many licensure and certification tests, in general), this is the appropriate design for the item bank, even though it yields different amounts of target information for the various pathways under the MST design.

Simulations

Seven simulation studies were conducted (see Table 2 for a summary of each of the studies). The first study used the final expected a posteriori (EAP) estimates from the medical certification test as the "true θ" to be estimated, and subsequently simulated responses based on these values. These data were selected so that the simulated exam θ distribution would resemble the operational certification data as closely as possible. The six additional studies all used random procedures to generate true examinee θs . Specifically, true examinee θs consisted of a random sample of values from a parametric θ distribution with specified parameters. Three of the studies incorporated true θs as a random sample from a normal distribution with mean of 1.00 and standard deviations of 0.75, 1.00, and 1.25, respectively. The remaining three studies incorporated true θs as a random sample from a beta distribution with parameters of (3, 2). These random deviates were subsequently transformed linearly to yield a mean of 1.00 and standard deviations of 0.75, 1.00, and 1.25, respectively. The linear transformation maintained the negatively-skewed shape of the distribution, though the means and the standard deviations were changed in accordance with the desired statistical moments. This particular beta distribution was selected so as to approximate the negative skew that is often observed with operational testing data (Lee, Brennan, & Kolen, 2006). The means and standard deviations for both the normal and the beta distributions were selected to approximate the certification test data as closely as possible.

Table 2. Mean and *SD* of θ for Three Simulation Distribution Conditions

Distribution	*N*	Mean	*SD*
"Real"	6,287	0.84	0.75
Beta			
	500	1.00	0.75
	500	1.00	1.00
	500	1.00	1.25
Normal			
	500	1.00	0.75
	500	1.00	1.00
	500	1.00	1.25

Note. Each simulation consisted of 100 replications

To begin each simulation procedure, each examinee was first administered the Stage 1 module consisting of 30 medium-level items. Item response strings (coded as 0/1 for an incorrect/correct response) were simulated for each examinee based on their true θ and the Stage 1 IRT parameters. Specifically, random uniform deviates were generated for each simulated examinee for each item; if the examinee's "true probability" of obtaining a correct response (calculated from "true θ" and the item parameters) was greater than the uniform deviate, the examinee was simulated as having correctly answered the item. Otherwise, the response was coded as incorrect. For each examinee, θ was estimated using the EAP scoring algorithm given the item responses and a standard normal prior distribution. EAP scoring was used rather than maximum likelihood estimation because the equation used in EAP scoring is closed form (it is not an iterative procedure) and EAP θ estimates can be obtained for all possible response patterns (which is not true in maximum likelihood estimation). For the linear design, examinees were then administered the Stages 2–6 medium modules. EAP scores were estimated after each stage.

For the MST design, routing rules were established based on the module-level TIFs. Specifically, the routing threshold was defined as the θ level where adjacent TIFs overlapped. For example, the easy-medium threshold was defined as the θ level where, to the left of this θ, the easy TIF yielded more information; to the right of this θ, the medium TIF yielded more information. Similarly, the medium-difficult threshold was defined as the θ level where, to the left of this θ, the medium TIF yielded more information; to the right of this θ, the difficult TIF yielded more information.

Following the Stage 1 module, EAP scores and corresponding standard errors of measurement (SEM; based on the square root of the Bayesian posterior variance) were calculated and the EAP scores were compared to the routing rules to determine Stage 2 modules for each examinee. Examinees were then administered either the Stage 2 easy, medium, or difficult module of items based on the routing rule. Following the Stage 2 administration, EAP scores and corresponding SEMs were recalculated for each examinee based on all previous items (i.e., Stage 1 and Stage 2 items) and Stage 3 modules were determined for each examinee based on the routing rules. This process continued until all six stages were administered. After Stage 6, final EAP scores and SEMs were estimated for each examinee based on all items administered to that examinee.

Along with calculating EAP scores after every stage, early termination indicators were also created for each examinee. Specifically, upper confidence limits were calculated for each examinee as the sum of the cut score and 1.96 multiplied by the estimated θ for each examinee. This follows the standard practice based on normal distribution theory for a 95% confidence limit. (Technically, this value is used for a two-tailed confidence limit, even though this was applied to asymmetric termination rules. The two-tailed confidence limit was retained, as early termination or different diagnostic routing rules may be used for extreme low performers.) Examinees were classified as "early termination" if the respective EAP estimate was above the confidence limit, signifying a statistically significant difference above the cut score.

Results

Estimation Accuracy

To investigate estimation accuracy, root mean squared error (RMSE) and mean SEM (MSEM) were calculated for each study and are presented in Table 3. RMSE provides an index reflecting the average squared difference between the estimated θ and the true θ across examinees. MSEM, on the other hand, provides an index of the confidence with which each exami-

nee's θ was estimated.

Table 3 shows that the MST procedure performed better than the linear procedure for each of the seven simulations on both of the evaluation statistics. That is, the MST procedure yielded consistently lower RMSE and MSEM values for each study. Furthermore, Table 3 reveals that both procedures performed better for distributions that were less variable (i.e., distributions with smaller standard deviations), and that both procedures performed more similarly when the distributions were less variable.

Table 3. RMSE and MSEM of θ Estimates and Their Difference (Diff.) for MST and Linear Tests, by Simulated Distribution

	θ		RMSE			MSEM		
Distribution	Mean	*SD*	MST	Linear	Diff.	MST	Linear	Diff.
"Real"	0.84	0.75	0.192	0.196	0.004	0.191	0.195	0.004
Beta								
	1.00	0.75	0.196	0.198	0.002	0.194	0.197	0.003
	1.00	1.00	0.204	0.210	0.006	0.200	0.206	0.006
	1.00	1.25	0.218	0.227	0.009	0.208	0.217	0.009
Normal								
	1.00	0.75	0.198	0.201	0.003	0.194	0.197	0.003
	1.00	1.00	0.210	0.216	0.006	0.200	0.205	0.005
	1.00	1.25	0.237	0.249	0.012	0.207	0.215	0.008

Whereas Table 3 provides overall indices of RMSE and MSEM across all replications, it was also of interest to determine how RMSE and MSEM compared across each replication. By comparing these statistics across each replication, it is possible to determine how well MST worked for small samples. That is, the overall indices provided an average of the RMSE and MSEM across all replications, and were based on a larger sample size; the comparison by replication determines these values based only on the sample within the particular replication, and therefore compares the MST and linear designs for small samples. Table 4 presents the percentage of replications for which MST outperformed the linear design on both criteria (RMSE and MSEM) for each of the simulation conditions. As expected, MST outperformed the linear design for the majority of replications. This percentage increased as the standard deviation of the distribution increased, which was also expected given the overall RMSE and MSEM results.

The fact that both procedures performed better for distributions with smaller dispersions is not unexpected. A well-known principle of IRT is that best measurement is obtained near the center of the score scale, where most of the item difficulties are located (Lord, 1980). Less precise measurement is obtained at either end of the scale, where there are fewer items that discriminate well. As a result, given that there are fewer examinees at the extreme ends of the scale when the standard deviation of the distribution is smaller, better measurement overall would also be expected given that most of the item difficulties are in the same region of the scale as the examinees. (Note that this is the same logic that drives the implementation of adaptive testing: measurement precision increases when item difficulties are located around examinee abilities.)

Table 4. Percent of Replications for Which MST Had Lower RMSE and MSEM

	θ			
Distribution	Mean	*SD*	RMSE	MSEM
"Real"	0.84	0.75	98	100
Beta				
	1.00	0.75	61	100
	1.00	1.00	79	100
	1.00	1.25	88	100
Normal				
	1.00	0.75	74	100
	1.00	1.00	73	100
	1.00	1.25	87	100

It might also be expected that MST would outperform the linear design to a greater degree for distributions with larger standard deviations. For the linear design, item difficulties for all three stages were centered around the middle of the θ scale. For the MST design, item difficulties in general were more spread out and examinees are administered sets of items that are more closely aligned with their estimated θ levels. Consequently, as the percentage of examinees scoring at the extreme ends of the scale increases (i.e., the standard deviation increases), more precise measurement overall would naturally be expected under the MST design in comparison to the linear design, given that the MST design accounts for measurement precision at either end of the scale as well as in the middle of the scale.

Although the RMSE and MSEM provide an indication of how well the MST and linear designs performed overall, it was also of interest to determine how well these designs performed at various regions along the IRT score scale. Figure 3 provides plots of RMSE by θ level for the MST and linear designs for each of the seven simulation studies and reveals that, as expected, measurement was much more accurate for θ levels near the center of the IRT scale. Furthermore, these figures reveal that MST outperformed the linear design almost consistently through the entire score scale. Additionally, these figures reveal that the greatest discrepancies between the MST and linear designs were in the lower half of the score scale.

Decision Accuracy and Decision Consistency

Whereas the intent of the previous analyses was to provide information concerning the accuracy with which θ was estimated, the accuracy with which pass-fail decisions are made (*decision accuracy*) and the consistency with which these decisions are made (*decision consistency*) are also very important considerations for licensure and certification tests. Tables 5 and 6 present decision accuracy and decision consistency estimates based on all replications for the study that incorporated the real certification test data.

Figure 3. RMSE Conditional on θ for the Seven Simulations

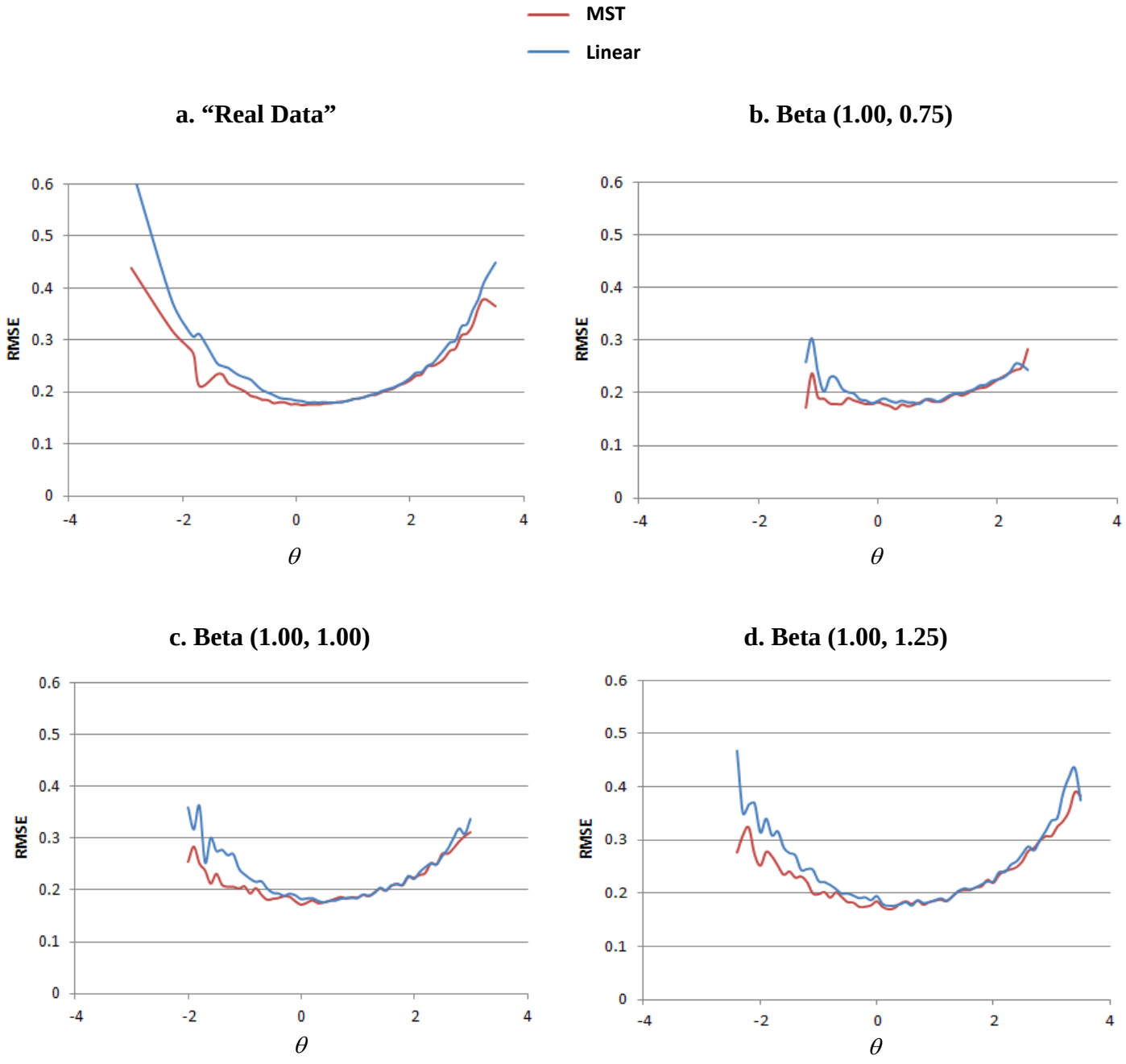


Figure 3 (continued). RMSE Conditional on θ for The Seven Simulations

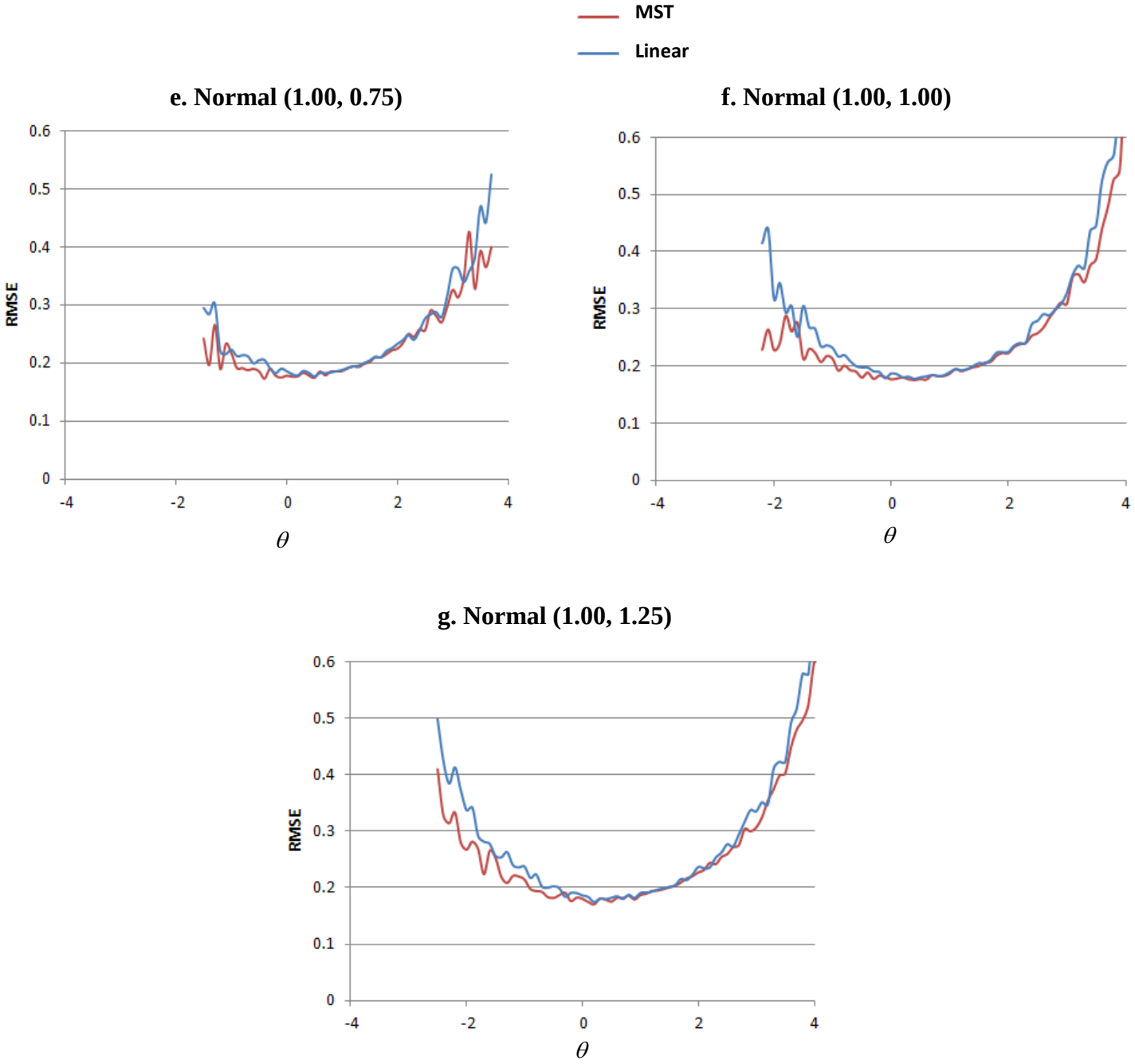


Table 5 reveals that the decision accuracy estimates for the MST design and the linear design were 94.41 and 94.26, respectively. This indicates that 94.41% of the examinees were correctly classified under the MST design (i.e., examinees with true θ above the cut score were classified above the cut score, and examines with true θ below the cut score were classified below the cut score) and that 94.26% of the examinees were correctly classified under the linear design. Although these results are not substantially different for these two designs, the MST design performed slightly better than the linear design according to this criterion.

Table 5. Percent Decision Accuracy for "Real" Test Data

	Estimated Decision					
	Linear			MST		
True Decision	Fail	Pass	Total	Fail	Pass	Total
Fail	16.34	2.33	18.67	16.33	2.34	18.67
Pass	3.41	77.92	81.33	3.24	78.08	81.33
Total	19.75	80.25		19.58	80.42	
Decision Accuracy			94.26			94.41

Table 6 presents the decision consistency estimates for the MST and linear designs. It should be noted that the decision consistency coefficients were calculated in a non-conventional fashion. For simulation studies that use only two replications, the conventional method for calculating decision consistency is to observe the percentage of observations that yield the same decision on both replications (i.e., the agreement percentage). Given that 100 replications were available, however, decision consistency was calculated differently.

Specifically, for each examinee, decision consistency was calculated as the percentage of replications that yielded a passing decision; this number was then subtracted from 1.0 if the value was less than 0.5. For example, an examinee whose estimated ability was above the cut score for each replication yielded a decision consistency value of 1.0; conversely, an examinee whose estimated ability was below the cut score for each replication yielded a decision consistency value of 1.0 − 0.0 = 1.0. This method was used so that consistent decisions—regardless of whether the examinee consistently passed or consistently failed the test—received decision consistency estimates of 1.0. Inconsistent decisions, on the other hand, were near 0.50. These examinee-level consistency estimates were then averaged across the 6,287 examinees to produce the statistics reported in Table 6.

Table 6 reveals that the decision consistency estimates for the MST and linear designs were 0.9447 and 0.9431, respectively. This implies that, not only were the pass-fail decisions slightly more *accurate* under the MST design, but the pass-fail decisions were slightly more *consistent* under the MST design as well.

Table 6. Decision Consistency for "Real" Test Data

Test Type	*N*	Mean	*SD*	Min	Max
Linear	6,287	0.9431	0.1170	0.50	1.00
MST	6,287	0.9447	0.1154	0.50	1.00

Early Termination

Another research objective was to investigate the early termination indices obtained under the MST design. That is, it was of interest to determine the percentage of examinees that would have terminated the test early if the early termination procedures had been enforced. Furthermore, along similar lines, it was of interest to determine the percentage of the examinees that would have terminated early in each of the stages. (Note that there is a statistical distinction between early termination and forced decision after the last stage, although there is no practical dis-

tinction. Early termination after the last stage implies that the EAP score is greater than—and significantly different than—the cut score. A forced decision after the last stage implies that the EAP score is greater than—but not significantly different than—the cut score).

Table 7 provides routing and pass patterns for the simulation consisting of all real certification examinees. Note that this particular simulation incorporated 100 replications of 6,287 examinees, and therefore pass-fail decisions were made for a total of 628,700 (100 × 6,287) examinees.

Table 7. Percentages of Examinees With Early Pass, and Percentage of Total Examinees Who Did Not Terminate Early by the Given Stage

	Path			Path		
Early Pass	Easy	Medium	Total	Easy	Medium	Total
Stage 1 (*N* = 628,700)				Stage 4 (*N* = 269,828)		
No	–	67.26	67.26	63.29	26.71	90.01
Yes	–	32.74	32.74	0.02	9.97	9.99
Total	–	100	100.00*	63.31	36.69	42.92*
Stage 2 (*N* = 422,863)				Stage 5 (*N* = 242,861)		
No	45.91	29.1	75.01	68.95	23.77	92.72
Yes	0.46	24.53	24.99	0.01	7.27	7.28
Total	46.37	53.63	67.26*	68.96	31.04	38.63*
Stage 3 (*N* = 317,195)				Stage 6 (*N* = 317,195)		
No	55.89	29.18	85.07	73.48	20.86	94.34
Ye s	0.09	14.85	14.93	0	5.66	5.66
Total	55.97	44.03	50.45*	73.49	26.51	35.82*
Forced Decision: No Early Termination (*N* = 212,443)						
No	57.67	0.09	57.76			
Yes	20.22	22.02	42.24			
Total	77.89	22.11	33.79*			

*Percentage of total examinees who did not terminate early by the given stage.

This table reveals that 628,700 examinees completed the Stage 1 medium module, and that 33% of these examinees (205,837) received an early termination status after Stage 1. Of the 422,863 examinees that did not terminate early after Stage 1, 46% (196,070) completed the Stage 2 easy module, and 54% (226,793) completed the Stage 2 medium module. It is interesting to note that all of the examinees who would have been assigned to complete the Stage 2 difficult module had already terminated early after the first stage. This is not unexpected, as examinees who are assigned to complete the most difficult module are the examinees at the highest θ levels.

Of the 422,863 examinees that completed either the Stage 2 easy or Stage 2 medium module, 25% (105,668) received an early termination status after the second stage. The remaining 317,195 examinees were all assigned to complete either the Stage 3 easy or medium module, which again implies that all of the examinees who would have been assigned to complete the Stage 3 difficult module had already terminated in the first or second stages. Similar patterns can be seen for Stages 4, 5, and 6. Ultimately, 122,710 examinees (19.5%) did not pass the test.

Discussion and Limitations

Comparison With Previous Studies

In general, the results obtained in this study are very similar to the results found in previous studies (Armstrong & Little, 2003; Guille et al., 2011; Hambleton & Xing, 2006; Jodoin et al., 2002, 2006; Luecht et al., 2006; Luecht & Burgin, 2003; Luecht & Sireci, 2011). For example, Jodoin, Zenisky, and Hambleton (2002, 2006) and Hambleton and Xing (2006) concluded that although the MST design did outperform the linear test design, the differences in decision consistency and decision accuracy for licensure and certification tests were not considerably dissimilar between the two designs. Furthermore, Guille et al. (2011) concluded that the MSEM was smaller under the MST design, which was also found to be true in this study. Whereas Guille et al. compared RMSE values for each of seventeen content areas—and concluded that RMSE was smaller under the MST design for twelve of the seventeen domains—the current study observed RMSE as aggregated across all domains and concluded that overall, the MST design outperformed the linear test design in regard to RMSE. Concerning early termination, it is difficult to compare the results found in this study with the results in previous studies, since early termination is based on many factors including the relationship between the cut score and the ability distribution, early termination criteria, and the exact MST design used, among other things.

Whereas the present study yielded comparable results to previous studies with regard to the overall measures of root mean squared error, mean standard error of measurement, decision accuracy, and decision consistency, the current investigations went beyond and expanded previous literature by comparing the MST and linear test designs for medium-sized tests. To do this, RMSE and MSEM were calculated and compared for *each* of the 100 medium-sized replications. The results revealed that the MST design slightly outperformed the linear design *for each replication* with regard to MSEM, and that the MST design typically outperformed the linear design with regard to RMSE across replications (Table 4). The study showed that these results also hold for medium-sized samples. Differences, however, were small.

Both procedures performed better when the ability distribution had a smaller standard deviation, and the MST design outperformed the linear design to an even greater extent when the ability distribution had a larger standard deviation. The fact that both procedures performed better for distributions with smaller dispersions is not unexpected. For distributions with smaller dispersions, relatively more examinees are located near the center of the scale, which results in fewer examinees completing the easy and difficult modules under the MST design. Therefore, the linear and MST designs yield more comparable results for distributions with smaller dispersions.

Along similar lines, the MST design might also be expected to outperform the linear design to an even greater extent for distributions with a larger dispersion. IRT scoring methods are known to produce more accurate and reliable ability estimates near the center of the scale, where most of the items are located (Lord, 1980). However, the MST is specifically designed to accommodate examinees performing at either end of the score scale, given that items are specifically targeted to these extreme ends. As a result, for distributions with greater dispersions (and therefore a greater percentage of examinees at either end of the score scale), it might be expected that the MST design would outperform the linear design to a greater extent.

The MST design also produced slightly higher decision accuracy and decision consistency indices across each of the seven simulation conditions, which might be expected given that the MST design yielded more precise ability estimates. The decision accuracy and decision consistency indices imply that not only were the pass-fail decisions more accurate under the MST design, but the pass-fail decisions were also more consistent under the MST design.

Limitations and Practical Implications for Item Bank Development

There are several limitations to this study due to the fact that real operational (and not simulated) conditions and item parameters were used to conduct the investigations (although response strings were simulated). Although these limitations do somewhat constrain the generalizability of these results, at the same time the limitations shed light on practical implications for operational testing programs that are considering implementing the MST design.

IRT discrimination parameters. The IRT discrimination parameter estimates for the operational test tended to be low, as evidenced in Table 1. This certainly affected the target TIFs and the adaptive efficiency of the MST design. Although the discrimination parameters were not ideal for MST, they are representative of the types of items that might be found in licensure and certification testing, especially in medicine. That is, licensure and certification tests are often administered to highly homogeneous groups of examinees, which significantly impacts the ability of items to discriminate. Unlike educational testing, where every student is tested and examinee ability varies substantially, the population for this particular test was highly selective and comprised of post-residency medical school graduates. As a result, the low discrimination parameters are partly attributable to the homogeneous population to which this test was administered.

It might be possible to obtain more discriminating items through targeted item writing, however. For example, through the use of automatic item generation (Gierl & Haladyna, 2013) and evidence-centered design models (Luecht, 2013), it might be possible to target item difficulties to specific locations on the score scale. This would not only change the shape of the item bank information function, in effect making it more accommodating for the MST design by distributing items along the score scale (recall that items for licensure-certification tests are often targeted at the cut score), but it might help to increase item discrimination, as more conscientious effort is being made to create items with pre-specified difficulty and discrimination properties.

Efficient use of test modules. It became apparent after the results were collected that under the early termination rules, no examinees completed any of the Stage 2–6 difficult modules (Table 7); all examinees who would have been administered these modules had already terminated at the end of Stage 1. As a result, the MST design used in this study (Figure 1) does not appear to be the most efficient design when asymmetric early termination is used for this particular test and—to generalize—when asymmetric early termination is used for tests with low or high cut scores relative to the ability distribution. There are several ways that testing programs in similar situations can proceed.

One method would be to have only two pathways rather than three. For example, rather than employing a 1–3–3–3–3–3 MST design in which the second through sixth stages are comprised of easy, medium, and difficult modules, it might be more efficient to employ a 1–2–2–2–2–2 design in which the second through sixth stages are comprised of only easy and medium modules.

An alternate method would be to create a full set of modules for the easy and medium pathways (as in Figure 1), but to create only a few modules for the difficult pathway. For example, whereas the easy and medium pathways would contain the full set of five modules following the initial medium-level module, the difficult pathway could contain only one or two modules. The difficult modules could then be administered in any stage in the event that an examinee who did not terminate early would be routed down the difficult pathway. This design would make much more efficient use of the item bank, as fewer difficult modules would be required and therefore fewer items would be susceptible to item exposure.

Although the two methods described above could be used to create a more efficient design within the MST framework, operationally speaking it might be difficult to use this in practice.

Specifically, it might be difficult to obtain the appropriate content balance if difficult modules are not administered in the test. This will vary by testing program, however, as it depends on the amount to which items vary in difficulty within content domains. For example, if one content domain is significantly more difficult than the other content domains on a particular test, it might be difficult to achieve appropriate content balance for this domain when using only easy and medium modules.

Ability estimation and stopping rules. EAP scoring was used to estimate θ in this study so that all examinees would receive θ estimates at the end of each stage (it is quite likely that some examinees had perfect scores at the end of the first stage, in which case maximum likelihood estimation would not have a solution). The prior distribution used for EAP scoring in this study was a standard normal distribution. The fact that EAP scoring was used in this study, along with the specific prior chosen for ability estimation, has implications for the results.

Considering that the "real" distribution had a mean of 0.84 and a standard deviation of 0.75 (all simulated distributions had a mean of 1.00), and that the cut score for this test was 0.165, the prior distribution in this study would shrink high scoring examinees toward the cut score. Conversely, very low performing examinees would also be regressed toward the cut score. If EAP scoring were to be used operationally for a testing program, the selection of a prior distribution becomes an important decision when the test is used to make pass-fail decisions; the prior distribution can shrink scores toward the cut score or push scores away from the cut score depending on the location of the cut score in comparison to the prior distribution.

On a related note, the conservative confidence value of 1.96, in conjunction with the EAP scores and their respective SEMs, had an impact on the early termination decisions. The choice of the two-tailed 95% confidence value is conservative in comparison to the less conservative one-tailed value of 1.65, which could have been used in this study. This certainly impacted the number of forced decisions required, as more examinees would have terminated the test early if a less conservative significance value had been used. Regardless of which significance value was used in the study, however, the examinees who did not terminate early would still have received the same EAP scores; in this case, the differentiation is primarily between whether each examinee passed the test with statistical confidence, or if the examinee passed the test based on a forced decision.

References

Armstrong, R. D., & Little, J. (2003, April). *The assembly of multiple form structures*. Paper presented at the annual meeting of the National Council on Measurement in Education.

Drasgow, F., Luecht, R. M., & Bennett, R. E. (2006). Technology and Testing. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 471–516). Washington, DC: American Council on Education.

Gierl, M. J., & Haladyna, T. M. (2013). *Automatic item generation*. New York, NY: Routledge.

Guille, R. A., Becker, K. A., Zhu, R. X., Zhang, Y., Song, H., & Sun, L. (2011, April). *Comparison of asymmetric early termination MST with linear testing*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

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. CrossRef

Jodoin, M. G., Zenisky, A., & Hambleton, R. (2002, April). *Comparison of the psychometric properties of several computer-based test designs for credentialing exams*. Paper presented at

the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Jodoin, M. G., Zenisky, A., & Hambleton, R. (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. CrossRef

Lee, W.-C., Brennan, R. L., & Kolen, M. J. (2006). Interval estimation for true raw and scale scores under the binomial error model. *Journal of Educational and Behavioral Statistics, 31*(3), 261–281. CrossRef

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

Luecht, R. (2013). An introduction to assessment engineering for automatic item generation. In M. J. Gierl & T. M. Haladyna (Eds.), *Automatic item generation* (pp. 59–76). New York, NY: Routledge.

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

Luecht, R., & Burgin, W. (2003). *Test information targeting strategies for adaptive multistage testing designs*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

Luecht, R., & Sireci, S. G. (2011). *A review of models for computer-based testing*. (Research Report). College Board. Retrieved from http://research.collegeboard.org/sites/default/files/publications/2012/7/researchreport-2011-12-review-models-for-computer-based-testing.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. CrossRef

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.

Acknowledgments

The authors would like to acknowledge Kirk Becker for providing invaluable suggestions for a more efficient MST design described in this paper. They would also like to acknowledge Renbang Zhu and Jane Zhang, who provided statistical advice and support for the data analyses.

Author Address

Bradley G. Brossman, American Board of Internal Medicine, 510 Walnut Street, Suite 1700, Philadelphia, PA 19106-3699. U.S.A. Email bbrossman@abim.org.