KURAM VE UYGULAMADA EĞİTİM BİLİMLERİ EDUCATIONAL SCIENCES: THEORY & PRACTICE

Received: September 20, 2016
Revision received: June 7, 2017
Accepted: July 6, 2017
OnlineFirst: August 21, 2017

www.estp.com.tr

DOI 10.12738/estp.2017.5.0484 ♦ October 2017 ♦ 17(5) ♦ 1759–1781

Research Article

Examining Content Control in Adaptive Tests: Computerized Adaptive Testing vs. Computerized Adaptive Multistage Testing

Halil İbrahim Sari[1]
Kilis 7 Aralik University

Anne Corinne Huggins-Manley[2]
University of Florida

Abstract

We conducted a simulation study to explore the precision of test outcomes across computerized adaptive testing (CAT) and computerized adaptive multistage testing (ca-MST) when the number of different content areas was varied across a variety of test lengths. We compared one CAT and two ca-MST designs (1-3 and 1-3-3 panel designs) across several manipulated conditions including total test length (24-item and 48-item test length) and number of controlled content areas. The five levels of the content area condition included zero (no content control), two, four, six and eight content area. We fully crossed all manipulated conditions within CAT and ca-MST with one another, and generated 4000 examinees from N (0,1). We fixed all other conditions such as IRT model, exposure rate across the CAT and ca-MSTs. Results indicated that test length and the type of test administration model impacted the outcomes more than the number of content area. The main finding was that regardless of any study condition, CAT outperformed the two ca-MSTs, and the two ca-MSTs were comparable. We discussed the results in connection to the control over test design, test content, cost effectiveness and item pool usage and provided recommendations for practitioner and also listed limitations for further research.

Keywords

Computerized adaptive testing • Computerized adaptive multistage testing • Content balancing

[1] **Correspondence to:** Halil İbrahim Sari (PhD), Muallim Rifat Faculty of Education, Kilis 7 Aralik University, Kilis 79100 Turkey. Email: hisari@kilis.edu.tr

[2] College of Education, University of Florida, 119A Norman Hall, PO Box 117049, Gainesville, FL, 32611, USA. Email: ahuggins@coe.ufl.edu

Citation: Sari, H. İ, & Huggins-Manley, A. C. (2017). Examining content control in adaptive tests: Computerized adaptive testing vs. computerized adaptive multistage testing. *Educational Sciences: Theory & Practice, 17,* 1759–1781. http://dx.doi.org/10.12738/estp.2017.5.0484

Introduction to the Adaptive Testing

The most widely used test administration model to measure student success today is the paper-and-pencil form. In this testing approach, the exam is administered on paper, the same set of items are given to all examinees, and item order cannot change during the test (e.g., American College Testing-ACT). One of the big advantages of paper-and-pencil testing models is high test developer control on content. This means that prior to the test administration, for each subject (e.g., biology) practitioners can specify related content areas (e.g., photosynthesis, ecology, plants, human anatomy, animals) and the number of items needed within each content area. However, drawbacks to paper-and-pencil tests are test security (e.g., cheating) which is a serious threat for test validity and score reliability (Thompson, 2008), low measurement efficiency, delayed scoring, late reporting and long test length (Yan, von Davier, & Lewis, 2014). Due to these deficiencies, adaptive testing has been proposed for use.

Another form of the test administration model is adaptive testing which is the main interest in this study. There are two versions of adaptive testing; computerized adaptive testing (CAT) and computerized adaptive multistage testing (ca-MST). In CAT (Weiss & Kingsbury, 1984), the test taker starts the exam with an item and then, depending on the performance on this item, the computer algorithm calculates his or her ability estimates, and selects the next question that contributes the most information about his or her current ability from the item pool. This process continues until the stopping rule is satisfied. The flowchart in Figure 1 visually displays working principle of CAT.

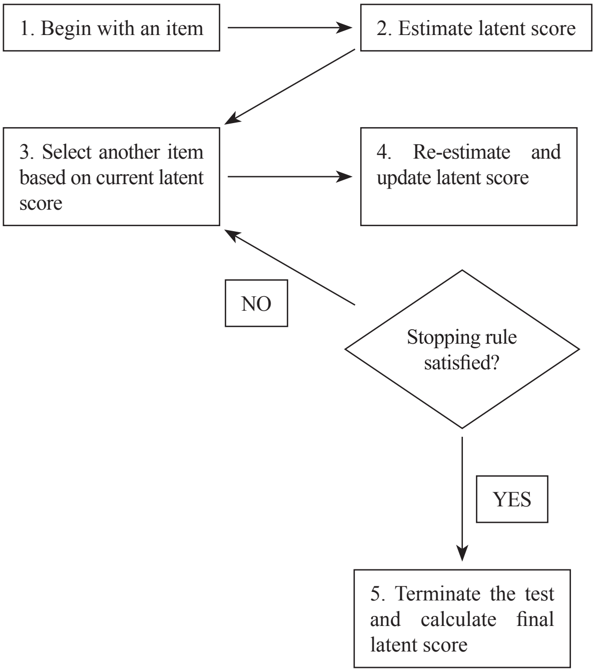
1. Begin with an item
2. Estimate latent score
3. Select another item based on current latent score
4. Re-estimate and update latent score
NO
Stopping rule satisfied?
YES
5. Terminate the test and calculate final latent score

Figure 1. The flowchart of CAT.

Second type of adaptive testing approach is computerized adaptive multistage testing (ca-MST). The ca-MST is comprised of different panels (e.g., a group of test forms), panels are comprised of different stages (e.g., division of a test), and stages are comprised of pre-constructed item sets at different difficulty levels called module (Luecht & Sireci, 2011). This means that at each stage some of the modules are easier and some of them are harder. In ca-MST, the test taker starts the test with a set of items (e.g., a set of 5 or 10 items) called routing module instead individual items. Depending on the performance on the routing module, the computer selects the next module in stage two that contributes the highest information about the test taker's current ability. This process continues until the test taker completes all stages. For illustration purposes, Figure 2 shows an example of ca-MST design with multiple panels. This design is called 1-3 ca-MST panel design, and there is one module in stage one, there are three modules in stage two. This two stage design is the simplest and widely used in both operational applications (e.g., revised version of GRE). This is because there is only one adaptation point in this configuration but this property also brings it to the disadvantage of higher likelihood of routing error (Yan et al., 2014). Armstrong, Jones, Koppel and Pashley (2004) and Patsula and Hambleton's (1999) huge simulation studies displayed that having more than four stages does not produce meaningful gain in test outcomes, and two or three stages with two or three modules at each stage are sufficient for a ca-MST administration (Armstrong, Jones, Koppel, & Pashley, 2004; Patsula & Hambleton, 1999; Yan et al., 2014). This is more likely due to having more adaptation points. Generating multiple panels is important to reduce panel, module, and item exposure rate, and prevents items from being overused. This is critical for test security; otherwise, test cheating and item sharing problems will arise.

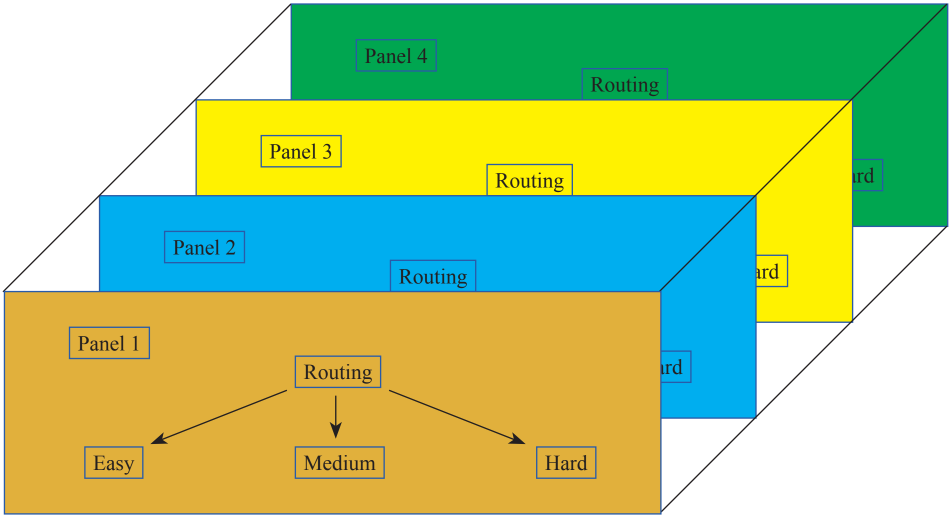

Figure 2. Illustration of multiple panels in ca-MST.

It is very clear that the main distinction between CAT and ca-MST is that there is an item level adaptation in CAT in contrast to the module level adaptation in ca-MST. The ca-MST has some advantages over other the CAT. Perhaps, the most obvious advantage of ca-MST over CAT is that ca-MST is more flexible in terms of item review and item skipping. The ca-MST allows examinees to go back to the previous items within each module, and to skip any item as well. However, examinees are not allowed to go back to the previous stage(s), and review items in the previous module(s).

Both CAT and ca-MST have multiple advantages over other administration models, such as immediate scoring, high measurement accuracy, low test length and high test security (Yan et al., 2014). However, one drawback to adaptive testing models is that unlike the paper-and-pencil tests it is not very easy to ensure all examinees are exposed to the same distribution of items in terms of content. This is extremely critical, especially in high stakes administrations (Huang, 1996), because unless the content distribution of items is the same across examinees, the test is essentially measuring different constructs for different persons.

Connection of Content Control with Test Fairness and Validity

Validity was once defined as a feature of a test device or instrument, but more recently the views and perceptions towards validity have changed (Kane, 2010). Messick's (1989) modern view on validity defines it as "integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment" (Messick, 1989, p.13). His well-accepted definition implies that validity is a broad issue, and in order to assess it, multiple source of evidence are required. This has been echoed in several chapters in the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council Measurement in Education [NCME], 1999). These sources include evidential and consequential basis evidences. Messick (1989) states that one must use these evidences for two purposes: a) interpretation of test scores (i.e., describing someone's test performance) and b) use of the test (i.e., making decisions based on the test performance). Although some still prefer the traditional taxonomy (e.g., construct validity, content validity, internal validity, external validity) (see Sireci, 2007), Messick argues validity is a unified concept, and that there is only one type of validity, which is construct validity, and all other types of validity are branched under it (Messick, 1989).

Construct validity refers to the degree to which a test instrument measures what it is supposed to measure (Cronbach, 1971). The essential goal of assessing construct validity is to cover all pieces of the construct of interest (e.g., critical thinking) as much as possible, and to show that there is no construct underrepresentation or overrepresentation. Construct overrepresentation occurs when a test's content measures more than what it intends to measure and that is irrelevant with the targeted construct of interest. Construct underrepresentation occurs when a test's content misses some important pieces of the targeted construct of interest. This is where content control and construct validity interact. For example, a test that measures high school students' general math ability must include test items from all related content areas as required by the use of and interpretation of the scores, such as algebra, fractions, arithmetic, story based problems, functions, calculus, trigonometry, geometry, and possibly more. If the test content lacks items from one or more of the required areas, the insufficient content coverage results in construct underrepresentation, and thereby negatively impacts construct validity. Consequently, this impacts the accuracy of measurement, interpretation of test scores, and thereby test fairness.

Purpose of the Study

As stated before, paper-and-pencil tests have a unique advantage of high test designer control. Thus, it is easier to ensure that all examinees receive a sufficient number of test items from all content areas. In fact, in terms of this feature, they are even incomparable with other test administration models. In contrast, there is a disadvantage in adaptive tests because in adaptive tests examinees receive overlapping or non-overlapping test forms. Without proper algorithms for content balancing, the test might be terminated for a test taker before receiving items from all content areas. Thus, it requires more consideration and effort to ensure that each examinee receives equal and sufficient number of items from all content areas. To be able to fully control test fairness and measurement accuracy, and to draw valid score interpretations, more consideration should be given to the issue of content balancing requirements in adaptive testing administrations.

Since adaptive tests have become popular, many statistical procedures (e.g., item selection rule, stopping rule, routing method) have been proposed and tested under a variety of conditions. Much research has been conducted on CAT including the proposal of new item selection methods (e.g., Barrada, Olea, Ponsoda, & Abad, 2008; Chang & Ying, 1996), stopping rules (e.g., Choi, Grady, & Dodd, 2010), and exposure control methods (e.g., Leung, Chang, & Hau, 2003; van der Linden & Chang, 2005). Researched areas of ca-MST include proposals for new routing methods (Luetch, 2000; Thissen & Mislevy 2000), test assembly methods (Luetch, 2000; Luecht & Nungester, 1998), and stage and module specifications (Patsula, 1999). Furthermore, many comparison studies have been conducted to explore efficiency of CAT versus ca-MST in terms of different test outcomes (see Davis & Dodd, 2003; Hambleton & Xing, 2006; Luecht, Nungester, & Hadadi, 1996; Patsula, 1999). However, the main focus in the comparison studies was the statistical components of CAT and/

or ca-MST, and little consideration has been given to the non-statistical aspects of adaptive tests such as content balancing control (Kingsbury & Zara, 1991). Since it is directly related to validity, score interpretation and test fairness, non-statistical issues of adaptive tests, such as content balancing, have not been given enough attention. The common findings in the literature was that due to the item-level adaptation and/or more adaptation points, CAT produces better accuracy of ability estimation (Yan et al., 2014). However, it is consistently asserted in several studies that a major advantage of ca-MST is that it controls for content better than CAT (see Chuah, Drasgow, & Luecht, 2006; Linn, Rock, & Cleary, 1969; Mead, 2006; Patsula & Hambleton, 1999; Stark & Chernyshenko, 2006; van der Linden & Glas, 2000; Weiss & Betz, 1974; Yan et al., 2014). Yet, the literature does not contain a study that specifically compares CAT with ca-MST under varying levels of content constraints to verify this claim. It is obvious that due to the feature of test assembly, ca-MST can easily meet content constraints. However, it is still unknown what we lose by administering ca-MST versus CAT when the two are different in how they select items from the same item bank to meet varying levels of content control goals. This study aims to explore the precision of test outcomes across the CAT and ca-MST when the number of different content areas is varied across a variety of test lengths. The goal of this study is to compare CAT and ca-MST in terms of content balancing control. This study aims to explore and compare the accuracy of outcomes produced by these two adaptive test approaches when strict content balancing is required. The study seeks answers to the following research questions:

1. How will the test outcomes be impacted when number of content area and test length are varied within each testing model (e.g., CAT, 1-3 ca-MST, 1-3-3 ca-MST)?

2. How will the test outcomes be impacted when number of content areas is varied under different panel designs (1-3 vs 1-3-3) and different test lengths (24-item and 48-item test length) on the ca-MST?

3. How will the test outcomes be impacted on the CAT and ca-MST under the combination of the levels of test length and content area?

Methodology

Design Overview

In this study, we compared one CAT design and two ca-MST panel designs across several manipulated conditions. We simulated two different ca-MST panel designs: the 1-3 and the 1-3-3 structure panel designs, with panels and modules constructed by integer programming. The common manipulated conditions across CAT and ca-MST were test length with two levels and number of controlled content area with five levels.

We fully crossed all manipulated conditions within CAT and ca-MST with one another. This resulted in 2x5 = 10 CAT conditions (test length x content area), and 2x5x2 = 20 ca-MST conditions (test length x content area x ca-MST design), for 30 total conditions. For each condition we performed 100 replications. For better comparability, we fixed the following features across CAT and ca- ca-MST administrations: exposure rate, item bank, item response theory (IRT) model, and ability estimation method. We detailed both varied and fixed conditions in following sections.

Fixed Conditions

The item parameters used in this study were based on a real ACT math test as used in Luecht (1991) and Armstrong, Jones, Li and Wu (1996). The original item bank consisted of 480 multiple choice items from six content areas. We provided the item parameters and number of items from each content in the original item bank in Table 1. In order to better compare the 30 conditions, and to avoid differences in content difficulty across the two, four, six and eight content area conditions, we generated four additional item banks (e.g., item bank 1, item bank 2, item bank 3 and item bank 4). We used these four different item banks in both CAT and ca-MST simulations, each for different content conditions. For the two content conditions, we selected one relatively easy (Content 1) and one relatively hard content area (Content 6) from six available on the real ACT test (i.e., item bank 1). Under the four, six, and eight content conditions, we used the same sets of ACT items as in the two content condition, with multiple modules being developed from those sets. All conditions had an equal number of easy and hard content areas within each item bank to avoid content difficulty problems in the interpretations of the results.

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

Content Area	*a*		*b*		*c*	
(Number of items)	Mean	SD	Mean	SD	Mean	
Content 1 (*n* = 48)	1.015	.292	-.485	.465	.154	.044
Content 2 (*n* = 168)	.911	.322	.131	.977	.160	.054
Content 3 (*n* = 24)	1.028	.328	.811	.778	.173	.059
Content 4 (*n* = 96)	1.120	.419	.689	.655	.167	.058
Content 5 (*n* = 96)	1.037	.356	.527	.650	.151	.062
Content 6 (*n* = 48)	.911	.312	.475	.828	.163	.058

All item banks had 480 items each with an equal number of items from each content area. There were 240, 120, 80 and 60 items from each content area in item bank 1, 2, 3 and 4 respectively. Under the no content control condition, we used item bank 1 but content ID's were ignored. We provided the item parameters and number of items for each content area in item bank 1 in Table 2. Again, the properties of items in other item pools are the same except the number of content areas and number of items. We provided the total information functions for the four item banks in Figure 3. As expected and desired, the level of average item bank difficulty was very similar across the item banks.

Table 2
Item Parameters of Each Content Area in Item Bank 1

Content Area	*a*		*b*		*c*	
(Number of items)	Mean	SD	Mean	SD	Mean	SD
Content 1 (=300)	1.015	.292	-.485	.465	.154	.044
Content 2 (=300)	.911	.312	.475	.828	.163	.058

In this study, we simulated many conditions for equivalency across CAT and ca-MST. These similarities allowed us to compare different content conditions not only within CAT and ca-MST designs, but also across the CAT and ca-MST designs. First, we generated four thousand examinees from a normal distribution, *N* (0, 1). We re-generated the theta values that represent examinees for each replication, and used for CAT and two ca-MST simulations. Second, under a particular content area condition, we built ca-MST modules and panels for two different panel designs from the same item banks used for CAT simulations. Specifically, in the two, four, six and eight content ca-MST conditions, when building the modules and panels for both 1-3 and 1-3-3 panel designs, we used item bank 1, 2, 3 and 4, respectively. In both ca-MST and CAT, we used the 3PL IRT model (Birnbaum, 1968) to generate the item responses. Third, we set exposure rates equal across the CAT and ca-MST simulations. The maximum item level exposure rate in CAT simulations was 0.25. The four essentially parallel panels created in ca-MST designs also had an exposure rate of 0.25 is for a panel and for a module. In practice, only routing modules are seen by all examinees that are assigned to a panel, and the subsequent modules in each panel are seen by fewer examinees. Fourth, as the ability estimation method, we used the expected a posteriori (EAP) (Bock & Mislevy, 1982) with a prior distribution of *N* (0, 1) for both interim and final ability estimates across the CAT and ca-MST. We

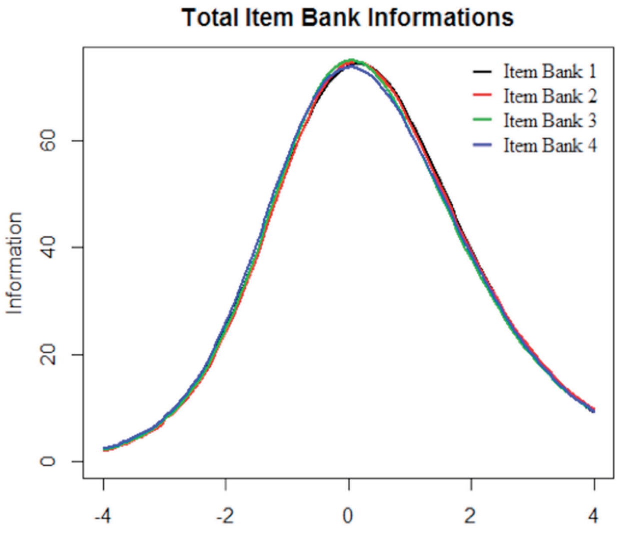

Figure 3. Total information functions for four item Banks.

completed the whole simulation process for both CAT and ca-MST in R version 2.1.1 (R Development Core Team, 2013). We used catR (Magis & Raiche, 2012) package for CAT simulations and wrote our own R code for ca-MST simulations. We detailed both varied conditions in CAT and ca-MST simulations in the following sections.

CAT Simulations

There were two varying conditions in the CAT design; test length with two levels and number of controlled content area with five levels. The two different levels of test length were 24-item and 48-item length as used in similar simulation studies (Zenisky, Hambleton, & Luecht, 2010). The five levels of content area condition included zero (e.g., no content control), two, four, six and eight content area. No content control means that students did not necessarily receive pre-specified number of items from each content area. For example, in a biology test, while one student might receive eight botany items, eight ecology items, eight organ systems items, another student might receive less balanced items from these content areas. In the two content condition, the target proportions for content 1 and 2 were 50% and 50%, and the corresponding number of items were 12, 12 and 24, 24 for 24-item and 48-item test length conditions, respectively. In the four content condition, the target proportions for content 1, 2, 3 and 4 were 25% each, and the corresponding number of items were 6 and 12 for 24-item and 48-item test length conditions, respectively. In the six content condition, the target proportions for content 1, 2, 3, 4, 5 and 6 were 16.6% each, and the corresponding number of items were 4 and 8 for 24-item and 48-item test length conditions, respectively. In the eight content condition, the target proportions for content 1, 2, 3, 4, 5, 6, 7 and 8 were 12.5% each, and the corresponding number of items were 3 and 6 for 24-item and 48-item test length conditions, respectively. Under the no content control condition, students received a total of 24 and 48 items without ensuring number of items they received from each content area. We summarized the distribution of items across the contents areas and test lengths in CAT simulations in Table 3. The literature showed that in terms of measurement accuracy, the constrained-CAT (CCAT; Kingsbury & Zara, 1989) method produced very similar results with other content control methods (Leung et al., 2003). So, the CCAT method was used for content control procedure as also used in many real CAT applications (Bergstrom & Lunz, 1999; Kingsbury & Zara, 1991). This method tracks the proportion of administered items for all contents. Then, the next item is selected from the content area that has the lowest proportion of administered items (e.g., largest discrepancy from the target proportion). The Sympson-Hetter (Sympson & Hetter, 1985) method with a fixed value of $r_{\mathrm{i}} = 0.25$ was used for item exposure control.

Since the first theta estimate is calculated after responding to the first item, the initial theta is not known prior to the CAT administration. A typical approach to choosing the first item is to select an item of medium difficulty (i.e., $b = 0$), which was used in this study. We used the maximum information method (Brown & Weiss, 1977) as the item selection rule. This method selects the next item that provides the highest information about his or her current theta estimate by satisfying the content constraints.

Table 3
Distribution of Items Across the Contents Areas and Test Lengths in CAT

Test Length	Content Condition	Content Areas	Target Proportions	Corresponding Number of Items
24-item CAT	Two Content	C_1,C_2	50%,50%	12,12
	Four Content	C_1,C_2,C_3,C_4	25%,25%,25%25%	6,6,6,6
	Six Content	C_1,C_2,C_3,C_4,C_5,C_6	16.6%,16.6%,16.6%, 16.6%,16.6%,16.6%	4,4,4,4,4,4
	Eight Content	$C_1,C_2,C_3,C_4,C_5,C_6,C_7,C_8$	12.5%,12.5%,12.5%,12.5%,12.5%,12.5%,12.5%,12.5%	3,3,3,3,3,3,3,3
48-item CAT	Two Content	C_1,C_2	50%,50%	24,24
	Four Content	C_1,C_2,C_3,C_4	25%,25%,25%25%	12,12,12,12
	Six Content	C_1,C_2,C_3,C_4,C_5,C_6	16.6%,16.6%,16.6%, 16.6%,16.6%,16.6%	8,8,8,8,8,8
	Eight Content	$C_1,C_2,C_3,C_4,C_5,C_6,C_7,C_8$	12.5%,12.5%,12.5%,12.5%,12.5%,12.5%,12.5%,12.5%	6,6,6,6,6,6,6,6

Ca-MST Simulations

We built two different ca-MST designs, the 1-3 structure design (e.g., two stage test) and the 1-3-3 structure design (e.g., three stage test). In any particular content area condition, all modules at any stage had the same number of items. In the 1-3 panel design, there were 12 and 24 items per module under 24-item and 48-item test length conditions, respectively. In the 1-3-3 panel design, there were 8 and 16 items per module under 24-item and 48-item test length conditions, respectively. Table 4 and 5 shows the distribution of items across the content areas in each module for 24-item and 48-item test length conditions, respectively. When the content was not controlled, the number of items in each module was the same under the stated conditions but the proportions across the content areas given in Table 4 and Table 5 were not necessarily met.

Previous research has shown that there are only slight differences between routing methods (Weissman, Belov, & Armstrong, 2007). In order to maximize the similarities between CAT and ca-MST, we selected the maximum information method (Lord, 1980) as the routing strategy because it can be similarly applied in both types of test administrations.

Test assembly. In both 1-3 and 1-3-3 ca-MST designs, we generated four non-overlapping essentially parallel panels from item bank 1, 2, 3 and 4. Creating multiple panels in ca-MST aimed to hold the maximum panel, module, and item exposure rates comparable to the CAT simulations. We used IBM CPLEX (ILOG, Inc., 2006) to create panels and modules. First, we clustered items into different modules, then randomly assigned modules to the panels. As shown in Luecht (1998), the automated test assembly finds a solution to maximize the IRT information function at a fixed theta point. Let denote θ_0 is the fixed theta point, and suppose we want a total of 24-item in the test. We first define a binary decision variable, x_i, (e.g., $x_i = 0$ means item *i* is not selected from the item bank, $x_i = 1$ means item *i* is selected from the item bank). The information function we want maximize is;

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

where ξ_i represents the item parameters of item *i* (e.g., *a*, *b*, *c* parameters). Let's say we have two content areas (e.g., C_1 and C_2), and want to select an equal number of items (e.g., 12 items) from each content area. The automated test assembly is modeled to maximize

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

subject to

$$\sum_{i \in C1}^{N} x_i \geq 12 \qquad (3)$$

$$\sum_{i \in C2}^{N} x_i \geq 12 \qquad (4)$$

$$\sum_{i=1}^{N} x_i \geq 24 \qquad (5)$$

$$x_i \in (0,1), i = 1, \ldots .N \, , \quad (6)$$

which put constraints on C_1, C_2, the total test length, and the range of decision variables, respectively. The test assembly models under other conditions (e.g., 48-item test length, six content area) can be modeled similarly. When we did not control content balancing, we removed the constraints on the contents from the test assembly model.

In all conditions, the three fixed theta scores were as $\theta_1 = -1$, $\theta_2 = 0$, $\theta_3 = 1$, which represent the target information functions for easy, medium and hard modules, respectively. In both panel designs, we chose the items in routing modules from

Table 4
The Distribution of Items in Modules in Ca-MST Across the Content Areas under 24-Item Test Length

	Content Condition	Content Area	Routing Modules	Stage 2 Modules	Stage 3 Modules	Total
1-3 Panel Design	Two Content	C_1,C_2	6,6	6,6	-	24
	Four Content	C_1,C_2,C_3,C_4	3,3,3,3	3,3,3,3	-	24
	Six Content	C_1,C_2,C_3,C_4,C_5,C_6	2,2,2,2,2,2	2,2,2,2,2,2	-	24
	Eight Content	C_1,C_2,C_3,C_4,C_5,C_6,C_7,C_8	2,2,2,2,1,1,1,1	1,1,1,1,2,2,2,2	-	24
1-3-3 Panel Design	Two Content	C_1,C_2	4,4	4,4	4,4	24
	Four Content	C_1,C_2,C_3,C_4	2,2,2,2	2,2,2,2	2,2,2,2	24
	Six Content	C_1,C_2,C_3,C_4,C_5,C_6	2,2,1,1,1,1	1,1,2,2,1,1	1,1,1,1,2,2	24
	Eight Content	C_1,C_2,C_3,C_4,C_5,C_6,C_7,C_8	1,1,1,1,1,1,1,1	1,1,1,1,1,1,1,1	1,1,1,1,1,1,1,1	24

medium difficulty items (e.g., items that maximize information function at theta point of 0). In the 1-3 panel design, there were two medium modules, one easy and one hard module in each panel. In the 1-3-3 panel design, there were two easy, three medium and two hard modules in each panel..

Table 5
The Distribution of Items in Modules In Ca-MSTs Across the Content Areas in under 48-Item Test Length

	Content Condition	Content Area	Routing Modules	Stage 2 Modules	Stage 3 Modules	Total
1-3 Panel Design	Two Content	C_1,C_2	12,12	12,12	-	48
	Four Content	C_1,C_2,C_3,C_4	6,6,6,6	6,6,6,6	-	48
	Six Content	C_1,C_2,C_3,C_4,C_5,C_6	4,4,4,4,4,4	4,4,4,4,4,4	-	48
	Eight Content	C_1,C_2,C_3,C_4,C_5,C_6,C_7,C_8	3,3,3,3,3,3,3,3	3,3,3,3,3,3,3,3	-	48
1-3-3 Panel Design	Two Content	C_1,C_2	8,8	8,8	8,8	48
	Four Content	C_1,C_2,C_3,C_4	4,4,4,4	4,4,4,4	4,4,4,4	48
	Six Content	C_1,C_2,C_3,C_4,C_5,C_6	4,4,2,2,2,2	2,2,4,4,2,2	2,2,2,2,4,4	48
	Eight Content	C_1,C_2,C_3,C_4,C_5,C_6,C_7,C_8	2,2,2,2,2,2,2,2	2,2,2,2,2,2,2,2	2,2,2,2,2,2,2,2	48

Evaluation Criteria

We evaluated the results of the simulation with two set of statistics: (a) overall results, (b) conditional result as evaluated in similar studies (see Han & Guo, 2014; Zenisky, 2004). For overall statistics, we computed mean bias, root mean squared error (RMSE), and correlation between estimated and true theta $(\rho_{\hat{\theta}\theta})$ from the simulation results as illustrated below.

Let N denote the total number of examinees, the estimated theta score for person *j*, and the true theta score for person *j*. Mean bias was computed as

$$\bar{e} = \frac{\sum_{j=1}^{N}(\hat{\theta}_j - \theta_j)}{N} \qquad (7)$$

The RMSE was computed as

$$RMSE = \sqrt{\frac{\sum_{j=1}^{N}(\hat{\theta}_j - \theta_j)^2}{N}} \qquad (8)$$

The correlation between estimated and true theta was computed as

$$\rho_{\hat{\theta}_j,\theta_j} = \frac{cov\,(\hat{\theta}_j,\theta_j)}{\sigma_{\theta_j}\sigma_{\theta_j}} \qquad (9)$$

where $\sigma_{\hat{\theta}_j}$ and $\sigma_{\hat{\theta}_j}$ are the standard deviations for the estimated and true theta values, respectively. In any particular condition in CAT and ca-MST simulations, we calculated each overall statistic separately for each iteration across the 4,000 examinees, and then averaged across 100 replications. In addition, we conducted factorial ANOVA procedures to examine the statistically significant and moderately to large sized patterns among the simulation factors on these three outcomes. We did this separately for each outcome with the three study factors (e.g., type of test administration- CAT vs 1-3 ca-MST vs 1-3-3 ca-MST, test length and number of content area) as fully crossed independent variables.

For conditional result, we calculated conditional standard error of measurement (CSEM) and observed between θ =-2 and θ=2, and the width of the θ interval was 0.2. CSEM for a theta point was computed as

$$CSEM(\hat{\theta}_j) = \frac{1}{\sqrt{T(\hat{\theta}_j)}} \qquad (10)$$

It should be noted that the CSEM is the inverse of the square root of total test information at an estimated theta point.

Results

The first section of the results describes the overall findings (e.g., mean bias, RMSE, correlations between the estimated and true theta values), and the second section describes the conditional result (e.g., CSEM).

Overall Results

Mean bias. We presented the results of mean bias across the conditions in Table 6. To assess for statistically significant patterns, we conducted a factorial ANOVA with mean bias as the outcome and the three study factors (e.g., test administration model, test length, number of content area) as the independent variables. The results for factorial ANOVA are in Table 7.

The main effect of test administration explained the largest proportion of mean bias variance ($\eta^2 = .17$), controlling for all other factors. The interactions or other main effects were either non-significant or explained very small proportion of variance. We displayed a graphical depiction of the main effect of test administration within each level of test length on the mean bias in Figure 4. The main finding was that regardless of the number of content area and test length, CAT produced lower amount of mean bias than the two ca-MTS's, which caused the main effect of test

Table 6
Results of Mean Bias Across the Conditions

Design	Test Length	No Control	2 Content	4 Content	6 Content	8 Content
1-3 ca-MST	24-item	0.07	0.07	0.05	0.05	0.06
1-3 ca-MST	48-item	0.06	0.07	0.06	0.07	0.07
1-3-3 ca-MST	24-item	0.06	0.05	0.06	0.05	0.07
1-3-3 ca-MST	48-item	0.08	0.05	0.05	0.04	0.06
CAT	24-item	0.009	0.009	0.009	0.009	-0.008
CAT	48-item	-0.008	0.007	-0.008	0.008	0.007

administration. Another finding was that the 1-3 and the 1-3-3 panel designs resulted in very similar mean bias. However, the test length did not impact the mean bias across the conditions.

Table 7
Factorial ANOVA Findings When Dependent Variable Was Mean Bias

Dependent Variable	*p*	η^2
Test Administration x Number of Content x Test Length	.15	.00
Test Administration x Number of Content	.04	.00
Test Administration x Test Length	.22	.00
Number of Content x Test Length	.25	.00
Test Administration	.00	.17
Number of Content Areas	.05	.00
Test Length	.61	.00

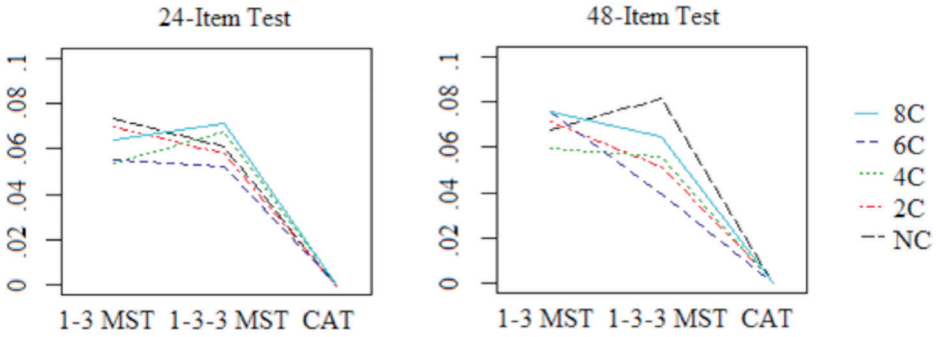

Figure 4. Main effect of Test Administration Model on Mean Bias within Levels of Test Lenghth.

Root mean square error. We presented the results of RMSE across the conditions in Table 8. To assess for statistically significant patterns, we conducted a factorial ANOVA with RMSE as the outcome and the three study factors (e.g., test administration model, test length, number of content area) as the independent variables. The results for factorial ANOVA are in Table 9.

The interaction of test administration model and test length explained a meaningful proportion of RMSE variance ($\eta^2 = .05$), as did the main effect of test length ($\eta^2 = .18$). We displayed a graphical depiction of the interaction of test administration model and test length within each level of number of content area on the RMSE in Figure 5. The main finding was that regardless of the number of content area and test administration model, as the test length increased, the amount of RMSE decreased. Also, the decrease in RMSE associated with the increase in test length was almost always more obvious for CAT, which caused the significant two way interaction of test administration and test length. The varying ca-MST panel designs did not impact the amount of RMSE within each level of test length as well as the number of content area.

Table 8
Results of RMSE Across the Conditions

Design	Test Length	No Control	2 Content	4 Content	6 Content	8 Content
1-3 ca-MST	24-item	0.33	0.33	0.35	0.33	0.34
1-3 ca-MST	48-item	0.30	0.30	0.31	0.31	0.30
1-3-3 ca-MST	24-item	0.32	0.32	0.33	0.34	0.34
1-3-3 ca-MST	48-item	0.30	0.31	0.31	0.31	0.31
CAT	24-item	0.35	0.36	0.36	0.37	0.36
CAT	48-item	0.28	0.29	0.29	0.28	0.29

Table 9
Factorial ANOVA Findings When Dependent Variable Was RMSE

Dependent Variable	*p*	η^2
Test Administration x Number of Content x Test Length	.04	.00
Test Administration x Number of Content	.48	.00
Test Administration x Test Length	.00	.05
Number of Content x Test Length	.38	.00
Test Administration	.06	.00
Number of Content Areas	.00	.02
Test Length	.00	.18

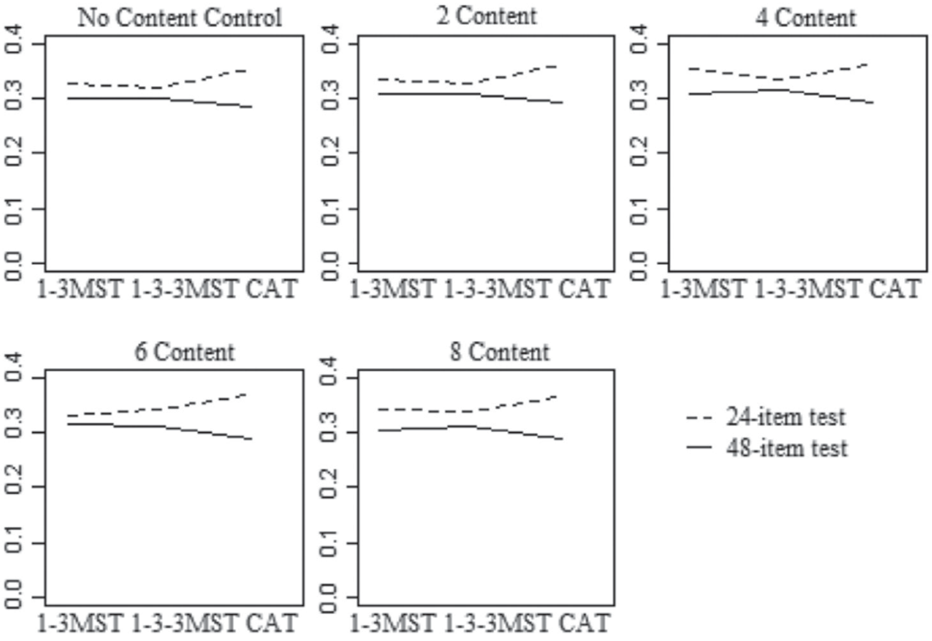

Figure 5. Interaction of Test Administration Model and Test Length on RMSE within Levels of Number of Content.

Correlation. We presented the results of correlation between the true and estimated theta values across the conditions were in Table 10. To assess for statistically significant patterns, we conducted a factorial ANOVA with correlation as the outcome and the three study factors (e.g., test administration model, test length, number of content area) as the independent variables. The results for factorial ANOVA are in Table 11.

The main effect of test administration and test length explained the largest proportion of correlation variance (η^2=.35), controlling for all other factors. The interactions or other main effects were either non-significant or explained very small proportion of variance. Figure 6 displays a graphical depiction of the main effect of test administration model within each level of test length on the correlations. The main finding was that regardless of the number of content area and test administration model, as the test length increased, correlation between the true and estimated thetas increased, which caused significant main effect of test length. Another main finding was that regardless of the number of content area and test length, the correlations were always lower under CAT, which caused significant main effect of test administration model. However, the number of content area did not impact the correlations. The varying ca-MST panel designs did not impact the correlations within each level of test length as well.

Table 10
Correlation Coefficients between True and Estimated Thetas Across the Conditions

Design	Test Length	No Control	2 Content	4 Content	6 Content	8 Content
1-3 ca-MST	24-item	0.95	0.95	0.95	0.95	0.95
1-3 ca-MST	48-item	0.97	0.97	0.97	0.97	0.97
1-3-3 ca-MST	24-item	0.96	0.95	0.95	0.95	0.95
1-3-3 ca-MST	48-item	0.97	0.97	0.97	0.97	0.97
CAT	24-item	0.93	0.93	0.93	0.92	0.93
CAT	48-item	0.95	0.95	0.95	0.95	0.95

Table 11
Factorial ANOVA Findings When Dependent Variable Was Correlation

Dependent Variable	*p*	η^2
Test Administration x Number of Content x Test Length	.12	.00
Test Administration x Number of Content	.04	.00
Test Administration x Test Length	.00	.02
Number of Content x Test Length	.00	.00
Test Administration	.00	.35
Number of Content Areas	.00	.00
Test Length	.00	.35

Conditional Result

Conditional standard error of measurement. We displayed the results of standard error of measurement conditioned on the estimated theta values across the two different test lengths and three different test designs in Figure 7. First finding was that standard error of measurements were always lowest around $\theta = 0$ point within any condition. Second finding was that as the test length increased standard error of measurements throughout the estimated theta points decreased for all test administration models. Third finding was that standard error of measurements for the two ca-MST conditions were more stable (more consistent) than CAT, regardless of the test length. Fourth finding was that even though the fluctuations were greater for

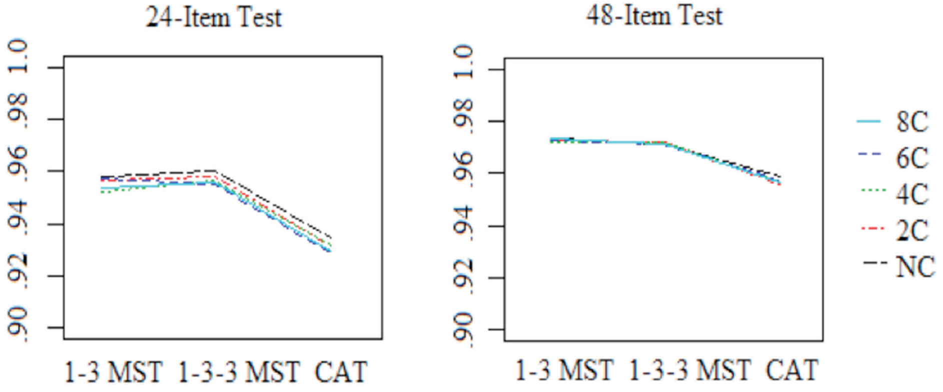

Figure 6. Main effect of test administration model on correlation within levels of test length.

CAT, CAT resulted in lower standard error of measurements throughout the estimated theta points. However, the number of content area did not substantially impact the standard error of measurements within a particular condition, and the interpretations were always similar across the number of content area conditions (see Figure 7).

Discussion and Limitations

Large scale tests in educational and psychological measurement are constructed to measure student ability using items from different content areas related to the construct of interest. This is a requirement for a validity because the degree of validity is related to showing that test content aligns with the purpose of the use of the test and any decisions based on test scores (Luecht, de Champlain, & Nungester, 1998). This means that a lack of test content from one or more areas severely impacts validity. It is well known that even if score precision is very high as would be the case through proper adaptive algorithms in CAT or ca-MST, this does not necessarily ensure that the test has valid uses (Crocker & Algina, 1986). For example, high score precision might not represent the intended construct if the content balancing in not ensured. Furthermore, if all students are not tested on all aspects of the construct of interest, test fairness is jeopardized. For these reasons, any test form adapted to an examinee has to cover all related and required sub-content areas for validity and fairness purposes.

The main purpose of this study was to explore the precision of test outcomes across computerized adaptive testing and computerized adaptive multistage testing when the number of different content areas was varied across the different test lengths. It was important to examine this because content balancing and content alignment is a requirement for validity of score-based inferences (Messick, 1989; Wise, Kingsbury, & Webb, 2015). In real applications, item pools most often have items from multiple content areas, and dealing with content control might not be easy in adaptive testing

(Wise et al., 2015). However, the consequences of not ensuring content balancing can have potential negative effects on the test use and score interpretations. Hence, this study added to the literature on content control in adaptive testing and, more specifically, aimed to provide guidelines about the relative strengths and weaknesses of ca-MST as compared to CAT with respect to content control.

The results showed that in terms of mean bias in theta estimates, CAT produced slightly better results than two ca-MSTs. This was the only meaningful finding, and other study factors did not have a substantial impact on the mean bias (see Table 6 and Figure 4). In terms of RMSE of theta estimates, only test length had a meaningful impact on the outcome (see Table 8); increasing test length improved the outcome (see Figure 5). In terms of correlations between the true and estimated theta values, both test length and type of test administration played an important role on the outcome (see Table 10). The two ca-MSTs produced very comparable results and outperformed CAT. Furthermore, increasing test length improved the correlation (see Figure 6). However, it is important to note that low correlations do not necessarily indicate that the results are poorer for CAT. As is seen in Figure 7, the instability in the conditional standard error of measurement was greater for CAT than for the two ca-MSTs. This was the reason behind the lower correlations between true and estimated theta under CAT. Even if the two ca-MSTs provided more stable standard error of measurements across the different theta values, CAT produced lower standard error of measurements than the two ca-MSTs (see Figure 7). The effect of test length was more obvious when the standard error of measurement was plotted against the theta values (see Figure 7).

This study did not find any evidence of the effect of number of content areas on both CAT and ca-MSTs. Increasing the number of controlled content areas or having no control over content did not meaningfully affect any of the study outcomes. All three test administration models were able to find items which provide the most information from different content areas, regardless of the number of content conditions. For practitioners and researchers, this is an indication that the studied CAT and ca-MST methods are not unduly influenced by the number of content areas one might have on a test. Thus, this is a positive finding for practitioners and researchers as CAT or ca-MST designers do not need to consider the number of content areas when concern is related to accuracy in theta estimation.

This study found that there was no meaningful difference on the outcomes between the two ca-MST panel designs, but CAT outperformed the ca-MSTs under many conditions. This study however does not argue against ca-MSTs. While the following considerations that we discuss are outside of the primary purpose of this study, it must be noted that ca-MSTs can have several practical advantages in operational testing. First, due to the features of test assembly, even if measurement precision is lower in ca-MSTs then CAT, all other parameters being equal, the ca-MST allows greater control over test design and content. The ca-MST designers can determine item and content order within modules. There is however no expert control over relative item order in CAT. Second, since items in the modules are placed by the test developer prior to the administration, ca-MST allows for strict adherence to content specification, no matter how complex it is (Yan et al., 2014). However, in CAT, content misspecifications are more likely to occur (see Leung et al., 2003). Third, ca-MST allows adding other constraints on placing items into the modules, such as item length and format. This means that the length of items in different modules and panels can easily be controlled in ca-MST. Fourth, ca-MST uses a higher percentage of an item pool as compared to CAT, all else being equal. In the 1-3 ca-MST condition, item pool usage rates were 40% and 80% under 24 and 48-item test lengths, respectively. In the 1-3-3 ca-MST condition, item pool usage rates were 46.6% and 93.3% under 24 and 48-item test lengths, respectively. However, pool usage rates for CAT were about 36% and 64% under 24 and 48-item test lengths, respectively. Fifth, ca-MST can have advantages for issues related to item retirement. The cost for a single item on a standardized test varies from $1,500 to $2,500 (Rudner, 2009). Having less retired items is desired because researchers and practitioners do not want to throw many items away after each administration. In ca-MST, only the items in the routing modules reach the pre-specified maximum control rate, and are then retired for use. In 1-3 ca-MST, the number of items with maximum exposure rates were 48 (e.g., 12 items per a routing module in a panel x 4 different panels) and 96 (e.g., 24 items per a routing module in a panel x 4 different panels) under 24 and 48-item length conditions, respectively. In 1-3-3 ca-MST, the number of items with maximum exposure rates were 32 (e.g., 8 items per a routing module in a panel x 4 different panels) and 64 (e.g., 16 items per a routing module in a panel x 4 different panels) under 24 and 48-item length conditions, respectively. In CAT, the number of items with maximum exposure rates were 47 and 105 under 24 and 48-item length conditions, respectively. Apparently, the 1-3-3 ca-MST resulted in the lowest number of items with maximum exposure rate. As a result, fewer items would have to be retired in operational practice. In fact, this number could be further reduced by placing less items into the routing modules, and thus more items can be saved for future administrations. This characteristic is another advantage of ca-MST over CAT. Further research can investigate item pool utilization under varying conditions such as item bank size, quality of item bank, different ca-MST panel designs and varying level of number of items in the routing module.

It is important to note that the content itself cannot be generated in a simulation study. Rather, item parameters are generated and used to represent different content areas. In this study, we defined a content area as a group of items, which belong to a specific sub-curriculum of the test, such as fractions, algebra or calculus. We justified this due to an alignment to operational parameter estimates from the ACT. However, when generating these content areas, we used a particular range of item parameters to represent such sub-curriculum sections of the test. This process has limitations, but is similar to approaches in other similar simulation studies (see Armstrong et al., 1996; Luecht, 1991).

This study argues that although ca-MST allows greater flexibility over the content ordering, there was a reduction in measurement efficiency. This is not surprising because of the less frequent number of adaptation points as compared to CAT. In ca-MST, the number of adaptation points is associated with the number of stages (e.g., one minus number of stages), whereas it is associated with number of items in CAT (e.g., one minus number of items). For example, in the 1-3 and 1-3-3 ca-MST panel designs, there was one and two adaptation points regardless of the test length, respectively. In CAT, there were 23 and 47 adaptation points under 24 and 48-item test length conditions, respectively.

In this study, we set the test length equal across the CAT and ca-MST simulations and investigated measurement accuracy as an outcome. However, measurement accuracy outcomes (e.g., standard error of measurement) are often used as stopping rules in CAT. In this study, ca-MST conditions were associated with a reduction in measurement accuracy as compared to CAT. However, if measurement accuracy were to be used as a stopping rule, we expect the outcomes to be quite different. It seems quite plausible that with this different type of stopping rule, CAT and ca-MST would have the same measurement accuracy outcomes but would differ in the test lengths needed to obtain those outcomes. In summary, the restriction of equal test length and the choice of stopping rule in this study had large impacts on the outcomes. Future studies may change and/or vary these conditions to explore outcomes across both test administration models.

In order to give similar advantages to both CAT and ca-MSTs, we intentionally used maximum Fisher information method as an item selection and a routing method. However, this study can be improved by running the same simulation while adopting other item selections and routing methods. Additionally, in order to avoid content difficulty confounding, we systematically chose two content areas as one easy and one hard content area. However, in real applications this will not likely happen and test content might have wider and non-systematic ranges of difficulty within contents areas. We recommend the studied conditions to be tested with a real test administration that does not have systematic difficulty differences across content areas. In a particular number of content conditions, there were equal number of items from different content areas. This study should also be replicated with unequal content distributions.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Armstrong, R. D., Jones, D. H., Koppel, N. B., & Pashley, P. J. (2004). Computerized adaptive testing with multiple-form structures. *Applied Psychological Measurement, 28*(3), 147–164.

Armstrong, R. D., Jones, D. H., Li, X., & Wu, L. (1996). A study of a network-flow algorithm and a noncorrecting algorithm for test assembly. *Applied Psychological Measurement*, *20*(1), 89–98.

Barrada, J. R., Olea, J., Ponsoda, V., & Abad, F. J. (2008). Incorporating randomness to the Fisher information for improving item exposure control in CATS. *British Journal of Mathematical and Statistical Psychology*, *61*, 493–513.

Bergstrom, B. A., & Lunz, M. E. (1999). CAT for certification and licensure. In F. Drasgow & J. B. Olson-Buchanan (Eds.), *Innovations in computerized assessment* (pp. 67–91). Mahwaj, NJ: Erlbaum.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Reading, MA: Addison-Wesley.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, *6*(4), 431–444.

Brown, J. M., & Weiss, D. J. (1977). *An adaptive testing strategy for achievement test batteries* (Research Rep. No. 77–6). Minneapolis, MN: University of Minnesota, Department of Psychology, Psychometric Methods Program, Computerized Adaptive Testing Laboratory.

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

Choi, S. W., Grady, M. W., & Dodd, B. G. (2010). A new stopping rule for computerized adaptive testing. *Educational and Psychological Measurement*, *70*(6), 1–17.

Chuah, S. C., Drasgow, F., & Luecht, R. (2006). How big is big enough? Sample size requirements for CAST item parameter estimation. *Applied Measurement in Education*, *19*(3), 241–255.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York, NY: Holt, Rinehart & Winston.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443–507). Washington, DC: American Council on Education.

Davis, L. L., & Dodd, B. G. (2003). Item exposure constraints for testlets in the verbal reasoning section of the MCAT. *Applied Psychological Measurement*, *27*(5), 335–356.

Hambleton, R. K., & Xing, D. (2006). Optimal and nonoptimal computer-based test designs for making pass–fail decisions. *Applied Measurement in Education*, *19*(3), 221–239.

Han, K. C. T., & Guo, F. (2014). Multistage testing by shaping modules on the fly. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing: Theory and applications* (pp. 119–133). Boca Raton, FL: Chapman and Hall/CRC.

Huang, S. X. (1996). A content-balanced adaptive testing algorithm for computer-based training systems. In C. Frasson, G. Gauthier, & A. Lesgold (Eds.), *Intelligent Tutoring Systems: Third International Conference Proceedings* (pp. 306–314). Heidelberg, Germany: Springer Berlin Heidelberg.

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

Kane, M. (2010). Validity and fairness. *Language Testing*, *27*(2), 177–182.

Kingsbury, G. G., & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education*, *2*(4), 359–375.

Kingsbury, G. G., & Zara, A. R. (1991). A comparison of procedures for content-sensitive item selection in computerized adaptive tests. *Applied Measurement in Education*, *4*(3), 241–261.

Leung, C. K., Chang, H. H., & Hau, K. T. (2003). Computerized adaptive testing: A comparison of three content balancing methods. *Journal of Technology, Learning, and Assessment*, *2*(5), 1–16.

Linn, R. L., Rock, D. A., & Cleary, T. A. (1969). The development and evaluation of several programmed testing methods. *Educational and Psychological Measurement*, *29*(1), 129–146.

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

Luecht, R. M. (1991). *American College Testing Program: Experimental item pool parameters* (Unpublished raw data).

Luecht, R. M. (1998). Computer-assisted test assembly using optimization heuristics. *Applied Psychological Measurement*, *22*(3), 224–236.

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

Luecht, R. M., de Champlain, A., & Nungester, R. J. (1998). Maintaining content validity in computerized adaptive testing. *Advances in Health Sciences Education*, *3*(1), 29–41.

Luecht, R. M., Nungester, R. J., & Hadadi, A. (1996, April). *Heuristic-based CAT: Balancing item information, content and exposure.* Paper presented at the annual meeting of the National Council of Measurement in Education, New York, NY.

Magis, D., & Raîche, G. (2012). Random generation of response patterns under computerized adaptive testing with the R package caR. *Journal of Statistical Software*, *48*(8), 1–31.

Mead, A. D. (2006). An introduction to multistage testing. *Applied Measurement in Education*, *19*(3), 185–187.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: American Council on Education/Macmillan.

Patsula, L. N. (1999). *A comparison of computerized adaptive testing and multistage testing* (Order No. 9950199). Available from ProQuest Dissertations & Theses Global (304514969).

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

R Development Core Team. (2013). *R: A language and environment for statistical computing, reference index* (Version 2.2.1). Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org

Rudner, L. M. (2009). Implementing the graduate management admission test computerized adaptive test. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 151–165). New York, NY: Springer.

Sireci, S. G. (2007). On validity theory and test validation. *Educational Researcher*, *36*(8), 477–481.

Stark, S., & Chernyshenko, O. S. (2006). Multistage Testing: Widely or narrowly applicable? *Applied Measurement in Education*, *19*(3), 257–260.

Sympson, J. B., & Hetter, R. D. (1985). Controlling item-exposure rates in computerized adaptive testing. *Proceedings of the 27th annual meeting of the Military Testing Association* (pp. 973–977). San Diego, CA: Navy Personnel Research and Development Center.

Thissen, D., & Mislevy, R. J. (2000). Testing algorithms. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (2nd ed., pp. 101–133). Hillsdale, NJ: Lawrence Erlbaum.

Thompson, N. A. (2008). A proposed framework of test administration methods. *Journal of Applied Testing Technology*, *9*(5), 1–17.

van der Linden, W. J., & Chang, H. H. (2005, August). *Implementing content constraints in alpha-stratified adaptive testing using a shadow test approach.* Law School Admission Council, Computerized Testing Report (01-09). Newtown, PA: Law School Admission Council.

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

Weiss, D. J., & Betz, N. E. (1974). *Simulation studies of two-stage ability testing* (No. RR-74-4). Minnesota, MA: Minnesota University Minneapolis Department of Psychology.

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

Weissman, A., Belov, D. I., & Armstrong, R. D. (2007). *Information-based versus number-correct routing in multistage classification tests* (Research Report RR-07-05). Newtown, PA: Law School Admissions Council.

Wise, S. L., Kingsbury, G. G., & Webb, N. L. (2015). Evaluating content alignment in computerized adaptive testing. *Educational Measurement: Issues and Practice*, *34*(4), 41–48.

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

Zenisky, A. L. (2004). *Evaluating the effects of several multi-stage testing design variables on selected psychometric outcomes for certification and licensure assessment* (Order No. 3136800).

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