ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2016; 7(2);388-406

Computer Adaptive Multistage Testing: Practical Issues, Challenges and Principles*

Bireye Uyarlanmış Çok Aşamalı Testler: Pratik Konular, Zorluklar ve Prensipler

Halil Ibrahim SARI** Hasibe YAHSI SARI *** Anne Corinne HUGGINS MANLEY****

Abstract

The purpose of many test in the educational and psychological measurement is to measure test takers' latent trait scores from responses given to a set of items. Over the years, this has been done by traditional methods (paper and pencil tests). However, compared to other test administration models (e.g., adaptive testing), traditional methods are extensively criticized in terms of producing low measurement accuracy and long test length. Adaptive testing has been proposed to overcome these problems. There are two popular adaptive testing approaches. These are computerized adaptive testing (CAT) and computer adaptive multistage testing (ca-MST). The former is a well-known approach that has been predominantly used in this field. We believe that researchers and practitioners are fairly familiar with many aspects of CAT because it has more than a hundred years of history. However, the same thing is not true for the latter one. Since ca-MST is relatively new, many researchers are not familiar with features of it. The purpose of this study is to closely examine the characteristics of ca-MST, including its working principle, the adaptation procedure called the routing method, test assembly, and scoring, and provide an overview to researchers, with the aim of drawing researchers' attention to ca-MST and encouraging them to contribute to the research in this area. The books, software and future work for ca-MST are also discussed.

Keywords: adaptive testing, computerized adaptive testing, computer adaptive multistage testing

Öz

Genel olarak eğitimdeki testlerin amacı testteki sorulara verilen cevaplarla testi alan bireylerin yetenek seviyelerini ölçmektir. Bu işlem yıllarca geleneksel yöntem olarak bilinen kağıt-kalem formundaki testlerle yapıldı. Ancak geleneksel yöntemler diğer test yöntemlerine (bireye uyarlanmış testler) göre yüksek ölçme hatası barındırmaları ve test uzunlugu gibi problemler nedeniyle çokça eleştirilmektedir. Bu problemlerin üstesinden gelebilmek icin bireye uyarlanmış testler tasarlanmıştır. Günümüzde kullanılan en yaygın bireye uyarlanmiş iki tip test bulunmaktadır: 1) bilgisayar ortamında madde bazında bireye uyarlanmış testler ve 2) bilgisayar ortamında modül bazında bireye uyarlanmış çok aşamalı testler. Madde bazında bireye uyarlamış testler yüzyılı aşkın bir geçmise sahip olup bugüne kadar üzerinde çokça çalışma yapılmıştır. Bu yüzden eğitimde ve psikolojide ölçme alanı dışındaki araştırmacılar tarafından bile birçok yönü itibariyle bilinmektedir. Fakat bireye uyarlanmış çok aşamalı testler, madde bazında bireye uyarlanan testlere göre çok daha yeni bir çalışma alanı. Bu sebeble de çok aşamalı testlerin birçok araştırmacı tarafından yeterince bilinmemektedir. Bu çalışmanın amacı bireye uyarlanmış çok aşamalı testlerin tüm özelliklerini, diğer testlerden farklılıklarını, avantajları ve dezavantajlarini araştırmacılarla paylaşmak, aynı zamanda araştırmacıların bu alana olan ilgilerini arttırmak ve bu alanın gelişmesine katki sağlamalarına teşvik etmektir. Çalışmada ayrıca bu alanda yazılan kitaplar, kullanılan bilgisayar yazılımları ve alanla ilgili gelecekte yapılabilecek çalışmalar tartışılmıştır.

Anahtar Kelimeler: madde ve modül bazında bireye uyarlanmiş testler, ölçme, psikometri

*This study is derived from a part of Ph.D. dissertation by Halil Ibrahim SARI entitled Content Control in Adaptive Tests: Computerized Adaptive Testing vs. Computerized Multistage Testing, submitted to University of Florida Graduate School under the supervision of Professor Anne Corinne HUGGINS-MANLEY.

**Post-Doctoral Associate, University of Florida, Department of Family, Youth and Community Services, Gainesville-Florida, United States of America, e-posta:hisari@ufl.edu
***Mathematics Teacher, Ministry of National Education, Yozgat-Turkey, e-posta:hsbyahsi@gmail.com
****Assistant Professor, University of Florida, Research and Evaluation Methodology Program, Gainesville-Florida, United States of America, e-posta:amanley@coe.ufl.edu

Received: 27.09.2016
DOI: 10.21031/epod.280183
Accepted: 27.11.2016

INTRODUCTION TO TEST ADMINISTRATION MODELS

There are numerous test administration models used to measure student achievement in the educational realm. Each model has advantages and disadvantages in terms of test validity, score reliability, test fairness, cost, practical issues, test administration, and schedule. The most widely used traditional model today is paper and pencil test. In this testing approach, the exam is administered on paper, the same set of items is given to all examinees, and item order cannot change during the test (e.g., the American College Testing). Since all examinees receive the same set of items, it is relatively easy to construct the forms because they do not always necessitate creating an item pool, which requires additional time, effort, and money. In addition, easier test administration, flexible item format (e.g., open-ended items), and item review by the examinees are primary advantages associated with paper and pencil tests (Becker & Bergstrom, 2013). However, one big criticism of paper and pencil tests is their vulnerability to security breaches (i.e., cheating). This is because all questions are exposed to all test takers, which is a serious threat for test validity and score reliability (Thompson, 2008). It is possible to see some examples of paper and pencil tests that overcome this challenge by creating multiple linear forms (e.g., the Scholastic Aptitude Test-SAT). The drawback to this testing model is delayed scoring and late reporting. This requires test takers to wait for their test scores to become available, which can present a problem if application deadlines are approaching. Long test length and low measurement efficiency can be counted as other major disadvantages of paper and pencil tests tests (Yan, von Davier, & Lewis, 2014).

The increased role of computers in educational and psychological measurement has led testing companies (e.g., Educational Testing Service [ETS], Pearson, KAPLAN, the College Board, and American College Testing [ACT]) and practitioners to explore alternative ways to deal with the deficiencies of linear tests. One solution is adaptive testing. The most widely known adaptive testing method is computerized adaptive testing (CAT) (Weiss, 1973). CAT has a long history in the field of educational measurement. In fact, the first attempt of computerized adaptive testing was intelligence tests created by Alfred Binet in 1905 (Wainer et al., 2000). It was also used during World War I for army recruitment purposes in the USA (DuBois, 1970). Throughout the past 100 years, much research has been conducted on CAT, including 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, 2002; van der Linden & Chang, 2005).

The working principle of CAT is as follows: First, computer algorithms randomly administer an item (typically an item of medium difficulty) to an examinee. After her response to the first item, the computer estimates her latent score and selects the next item from the pool that best matches with her current trait level. This basically means that when she gets an item correct, the computer selects a harder question; if she gets it wrong, the computer selects an easier question. This process continues until the stopping rule is satisfied. A flowchart in Figure 1 visually summarizes this. Even though it is successful in how accurate it measures ability and how secure it is—in contrast to linear testing—CAT does have its own disadvantages. First, it requires a large item pool which incurs a high cost to testing companies. Second, it requires complicated software and fast computers to be available in test centers. The third drawback is that CAT generally does not allow examinees to review the items that have already been answered or to skip any item during the test. Test takers usually have to respond to all items, and cannot go back to previous items.

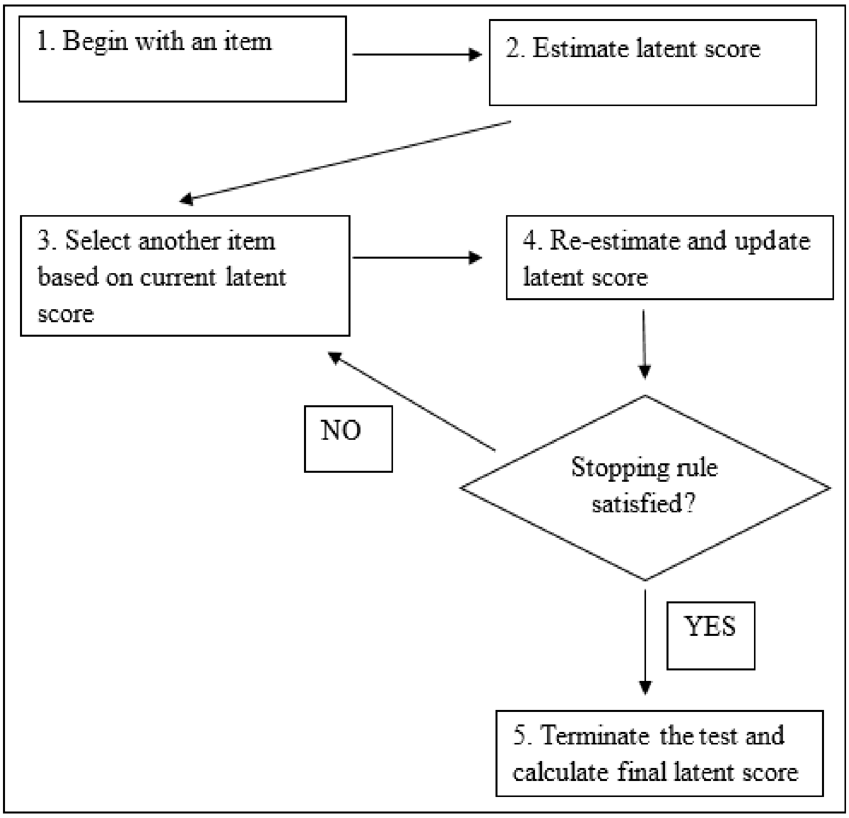
1. Begin with an item
2. Estimate latent score
3. Select another item based on current latent score
4. Re-estimate and update latent score
NO
Stopping rule satisfied?
YES
5. Terminate the test and calculate final latent score

Figure 1. The Flowchart of Computerized Adaptive Testing

To deal with some of the latter's deficiencies, multistage testing (MST) has been proposed as an alternative test administration method. Over the years, it has been called by different names, such as two-stage testing (Kim & Plake, 1993), computerized mastery testing (Lewis & Sheehan, 1990), computer adaptive sequential testing (Luecht, 2000), and bundled multistage testing (Luecht, 2003). Although multistage testing is relatively new compared to linear tests and CAT, the MST idea is not new (Yan et al., 2014). Early MST designs date back to the 1950s (see Angoff & Huddleston, 1958). The first versions of MSTs were in paper-and-pencil format, and there was no adaptation from one point to another (see Cronbach & Glaser, 1965; Lord, 1971 & 1974; Weiss, 1973). Even though these initial attempts were amateur compared to today, they were invaluable in terms of growing the field of alternative test administration.

The early 1970s were the most critical years for MST advancements. That is because Fredrick Lord and David Weiss created the basis of the first item response theory (IRT)-based MST applications. When the research on CAT eclipsed MST, an adaptive version of multistage testing, which is what this study is concerned with, was proposed for use (Keng, 2008; Mead, 2006). It is called computer adaptive multistage testing (Yan et al., 2014), or ca-MST, and it has gained in popularity recently, with more than one hundred journal articles published in the past twenty years alone. In addition, the number of operational examples has increased recently— tests using the ca-MST format include the Massachusetts Adult Proficiency Test, Graduate Record Examination or GRE, Law School Admission Council or LSAC, and Certified Public Accountants or CPA Examination. After the GRE switched formats from CAT to ca-MST in 2011, interest in ca-MST increased exponentially. But despite this growing interest in ca-MST, we believe that it has not yet received the recognition that it deserves.

Purpose of the Study

Due to its long history, CAT is a well-known test administration model. However, many features of ca-MST are still unknown by many researchers, especially outside the U.S. So this paper briefly reviews the ca-MST literature, compiling recent developments for researchers newly interested in ca-MST.

The purpose of this study is to closely examine the characteristics of ca-MST, including its working principle, the adaptation procedure called the routing method, test assembly, and scoring, and provide an overview to researchers and practitioners, with the aim of drawing researchers' attention to ca-MST and encouraging them to contribute to the research in this area. To that end, this study first explains ca-MST in detail, summarizes the recent developments in the literature of the area of ca-MST, and then discusses future work in ca-MST research.

STRUCTURE AND WORKING PRINCIPLE OF CA-MULTISTAGE TESTING

The ca-MST terminology includes some special terms not used in other testing procedures. These include module, stage, panel, routing, path, and test assembly. Ca-MST is made up of different panels (e.g., a group of test forms), and those panels are, in turn, composed of different stages (e.g., division of a test). The stages themselves are made up of pre-constructed item sets, called modules, at different difficulty levels (Luecht & Sireci, 2011). This means that at each stage some of the modules are easier and some of them are harder. In other words, some modules are more appropriate for low-ability test takers, while some are more appropriate for high-ability test takers.

There is almost always one module in stage one, called the routing module, which is used to establish the test taker's proficiency level. The test taker moves to the next module of the test based on her performance on the routing module. The number of stages, the number of the modules in each stage, and the number of the items in each module can vary from test to test.

In general, the working principle of a ca-MST is as follows. After assigning a test taker to a panel, unlike individual items in CAT, ca-MST starts with a routing module (e.g., a set of five or ten items). After the routing module, the first stage of ca-MST, the computer calculates the test taker's latent performance. Then, based on her current performance, the computer selects one of the pre-constructed modules in the second stage, and routes the test taker to the appropriate module. For example, if her performance on the routing module is high, she receives a harder module in the second stage; otherwise an easier module is selected. After she completes the second stage, again, the computer calculates her performance, and routes her to the most appropriate module in the third stage. This process continues until the test taker completes all stages. This is the main distinction between CAT and ca-MST: there is an item level adaptation in CAT, in contrast to the module level adaptation in ca-MST. This feature brings the advantages of item review, item skip, higher control over test content, strict adherence to the target content distributions, and consistent item order.

Figure 2 shows an example of the simplest possible ca-MST structure, with two stages. There is one module in stage one and two modules in stage two; this structure is called the 1-2 panel design. As shown in this figure, there are two possible pathways that a test taker might draw (e.g., Routing-Easy and Routing-Hard).

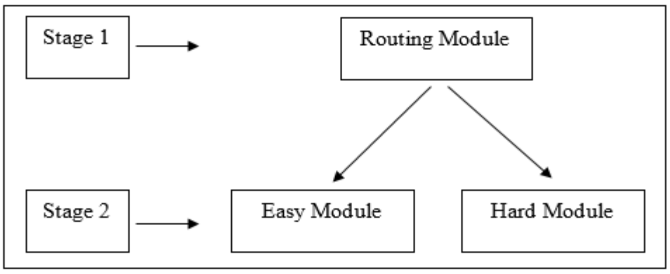

Figure 2. An Example of 1-2 Ca-MST Panel Design

Figure 3 shows a more complex ca-MST design, with three stages; there is one module in stage one and three modules in stages 2 and 3. Accordingly, this structure is called the 1-3-3 panel design. There are seven possible paths in this type of ca-MST design. These are: Routing-Easy-Easy, Routing-Easy-Medium, Routing-Medium-Easy, Routing-Medium-Medium, Routing-Medium-Hard, Routing-Hard-Medium, Routing-Hard-Hard. As can be understood from the figure, the pathways from a module to another module in the next stage that is not adjacent to the current module are ignored. For example, if a student receives easy module in stage 2, the student is not permitted to receive the hard module in stage 3, even if she performed very well in stage 2. The strategy of disallowing extreme jumps among the module is very common in many ca-MST designs (Luecht, Brumfield, & Breithaupt, 2006), and prevents aberrant item responses and inconsistent response patterns (Davis & Dodd, 2003; Yan et al., 2014). However, this is something that a test developer needs to decide prior to the test administration.

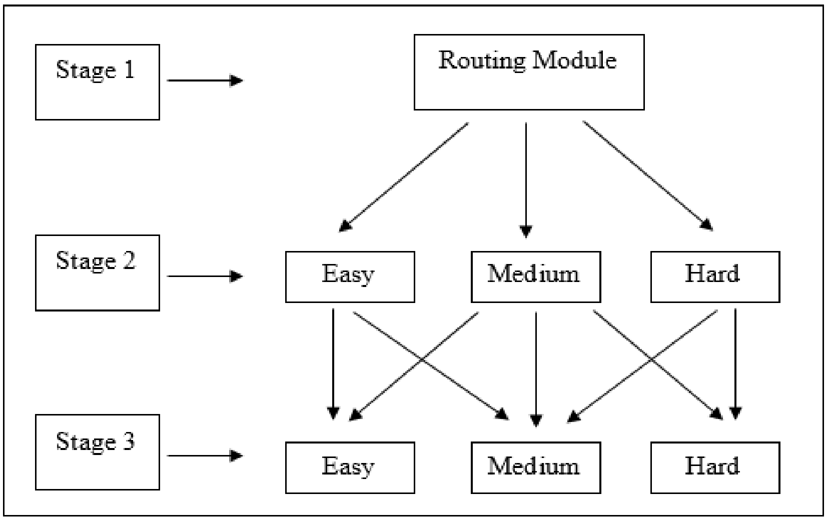

Figure 3. An Example of 1-3-3 Ca-MST Panel Design

Both Figures 2 and 3 display an example of one panel only, whereas a computer adaptive multistage panel is referred to as a collection of modules (Luecht & Nungester, 1998). In order to control panel exposure rate (and thereby module and item), ca-MST designs consist of multiple panels that are similar to each other, and each test taker is randomly assigned to one of them. Figure 4 shows an example of multiple panels. The panel construction procedure is described in the following sections.

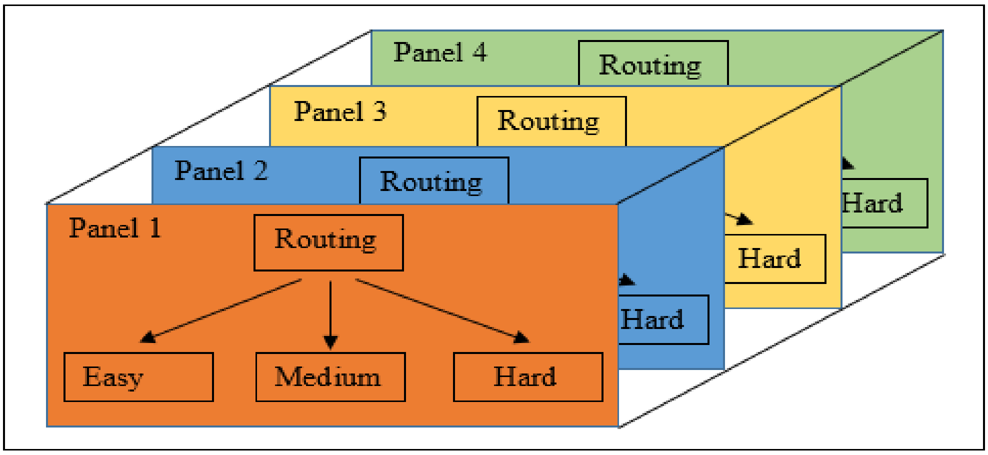

Figure 4. Illustration of Multiple Panels in Ca-MST

Like all other test administration models, ca-MST has advantages and disadvantages. Perhaps the most attractive advantage of ca-MST over CAT is that ca-MST is more flexible in terms of item review and item skipping. 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). Compared to linear tests, the ability to go back to a limited number of previous items is a drawback, but compared to traditional CAT, it is a remarkable feature. There is a well-known motto that the first answer is always correct, and students should always trust their initial responses (van der Linden, Jeon, & Ferrara, 2011). However, a research by Bridgeman (2012), an ETS practitioner, showed that this common belief is actually a superstition, finding an increase in student abilities when examinees had a chance to review answers (Bridgeman, 2012).

In terms of test length, ca-MST typically falls somewhere between linear tests and CAT (Hendrickson, 2007). Ca-MST is a fixed test, which means that test length is determined by the test developer; however, in CAT, the test length can be fixed or varied. In terms of test design control (e.g., content area, answer key balance, and word count), ca-MST allows more flexibility than CAT but less than linear tests. In terms of measuring ability, ca-MST is much more accurate than linear tests (Armstrong, Jones, Koppel, & Pashley, 2004; Patsula, 1999) but about as accurate as or slightly less accurate than CAT (Armstrong et al., 2004; Kim & Plake, 1993; Patsula, 1999).

However, some of CAT’s drawbacks, such as the high cost of creating an item bank, are still a concern in ca-MST. Also, while CAT can stop at any point (if the stopping rule is satisfied), reducing the test time, due to the module level adaptation, ca-MST continues until all stages are completed (Zheng, Nozawa, Gao, & Chang, 2012).

By taking into account the advantages and disadvantages of ca-MST, it can be said that ca-MST is a highly promising test administration model because it combines the advantages of linear tests and CAT. Ca-MST shares some characteristics with CAT, such as test design and structure, routing method, ability estimation, content control and test assembly, and exposure control. These components are summarized in later sections.

BUILDING A COMPUTER ADAPTIVE MULTISTAGE TEST

Prior to administering a ca-MST, the test designer must determine several things. These include the number of panels, stages, modules, items, and total number of test items, as well as automated test assembly (ATA), content control, routing method, and interim and final ability estimation method. These steps are summarized in the following sections.

Number of Panels

As stated before, parallel test forms called panels comprise the ca-MST, and the assignment of test takers to the panels occurs randomly (Luecht, 2003). Having multiple panels helps 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 (Yan et al., 2014). Depending on the importance of the exam (e.g., high stake or low stake), the number of panels changes, but in both operational examples and simulation studies it usually varies from one to forty (Yan et al., 2014). The preferred exposure rate in CAT generally ranges from 0.20 to .35. In order to achieve the similar exposure rate in the area of ca-MST—for example, if the desired panel exposure rate is .20 (as in the CAT version of the GRE)—one needs to create five ca-MST panels (the exposure rate is $1/r$, where r is the number of panels) (Eignor, Stocking, Way, & Steffen, 1993). Setting a low panel exposure rate means having more panels and more items.

We know that having a lower number of retired items is desired, because researchers and practitioners do not want to throw many items away after each administration. This is because item-writing is not easy work, and requires professionalism and money. A qualified item has to meet certain criteria, such as good discrimination and content appropriateness. Thus items have to be created by content professionals. Furthermore, the estimated cost for a single qualified item ranges between $1,500 and $2,500 (Rudner, 2009). So if we want to have a bank size of 500, it will cost between $750,000 and $1,250,000. In fact, this amount will increase, as the bank needs maintenance. However, in reality, only the items in the routing module are received by all test takers that are assigned to that panel, and they are then retired for use. Yet, due to the adaptation feature of ca-MST, the following modules and thereby items are received by fewer test takers. Thus, the desired panel exposure rate is the maximum exposure rate for a panel. Since some of the items in different panels do not reach the pre-specified exposure rate, they can be used in later administrations. In other words, the number of panels should be decided by taking into account a combination of desired test security and cost. It is important to note that a test designer can have overlapping panels, which means that some of the items can be placed in more than one panel. Also, so some of the items can be saved for future administrations of the test. The panel construction procedure is handled in the following sections.

Number of Stages and Modules

As stated before, the modules in each stage differ in their average difficulty levels, but the number of items and the proportion of content specifications are the same across the modules in a stage (i.e., they are not necessarily the same in modules at different stages). There are several examples of ca-MST design configurations in the literature such as 1-2 (e.g., Wang, Fluegge, & Luecht, 2012), 1-3 (e.g., Schnipke & Reese 1999; Wang, Fluegge, & Luecht, 2012), 1-2-2 (e.g., Zenisky, 2004), 1-3-3 (e.g., Keng & Dodd, 2009; Luecht et al., 2006; Zenisky 2004), 1-2-3 (e.g., Armstrong & Roussos 2005; Zenisky 2004), 1-3-2 (e.g., Zenisky,2004), 1-1-2-3 (e.g., Belov & Armstrong, 2008; Weissman, Belov, & Armstrong, 2007), 1-5-5-5-5 (e.g., Davey & Lee 2011), 1-1-2-3-3-4 (e.g., Armstrong et al. 2004), 5-5-5-5-5-5 (e.g., Crotts, Zenisky, & Sireci 2012). The two-stage design is the simplest and most widely used in both operational applications (e.g., the revised version of GRE) and simulation studies. This is because there is only one adaptation point in this configuration, but this same quality also has the disadvantage of introducing a higher likelihood of a routing error (Yan et al., 2014). Some have suggested using an additional module called a “recovery module” when necessary, but it was not found very interesting (see Schnipke & Reese, 1999). Thus, as previous studies showed, although the test complexity increases, adding more stages and modules into each stage and/or allowing more branching increases test outcomes (Luecht & Nungester 1998; Luecht, Nungester, & Hadadi 1996). This is likely due to having more adaptation points. Having more adaptation points makes ca-MST similar 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 in CAT, it is associated with the number of items (e.g., one minus number of items). For example, in a 1-3-3 ca-MST panel design, there are two adaptation points regardless of the test length. In a 20-item CAT, there are 19 adaptation points.

Some researchers believe that having multiple modules at the last stage is better for gaining accuracy in ability estimations, because the estimated abilities become closer to the test taker’s true abilities at the end of the test (Luecht & Nungester 1998). Armstrong et al. (2004) and Patsula and Hambleton’s (1999) huge simulation studies showed that having more than four stages does not produce meaningful gains in test outcomes, and two or three stages with two or three modules at each stage are sufficient for a successful ca-MST administration (Armstrong, et al., 2004; Patsula & Hambleton, 1999; Yan et al., 2014). The characteristics of modules are handled in the following sections.

Number of Items

Ca-MST is a fixed-length test, so it is necessary to decide the total test length for a test taker. The total test length varies from ca-MST to ca-MST (e.g., from 10 to 90). Then it is necessary to decide the number of items in each module. Based on the literature, tests with 20 or fewer items are usually considered short tests (Thompson & Weiss, 2011), while tests with 60 or more questions are generally considered long tests (Zheng et al, 2012).

After deciding the total test length, it is possible to assign a different number of items to the modules at different stages. The research on number of items can be summarized as follows: a) adding more items to tests increases the reliability of test scores and thereby measurement accuracy (Crocker & Algina, 1986); b) varying the number of items in modules at different stages does not affect test outcomes (Patsula & Hambleton, 1999); c) having shorter modules is better because it allows for more adaptation points; and d) increasing the number of items in the routing module has a positive effect on test outcomes (Kim & Plake, 1993). For this last item, Kim and Plake (1993) claimed having more items decreases the likelihood of misrouting in the following stages. However, as discussed before, having more items in the modules at the last stage can also have a positive effect on test outcomes. So, it can be concluded that when we earn from one side, lose from another side (Zheng et al, 2012).

Routing Method

The routing method in ca-MST is analogous to the item selection method in CAT. Since there is a module level adaptation in ca-MST, the routing method is used to decide the subsequent modules that an examinee will receive (Luetch, 2003). The efficiency of the routing method affects the pathway a student draws during the test, so misrouting impacts the test outcomes, and therefore the usefulness of ca-MST (Lord, 1980). Thus, it is a tremendously important component in ca-MST. Routing methods are generally sorted into two categories, dynamic rules and static rules (Yan et al., 2014). The most widely known dynamic method of maximum Fisher's information (MFI) (Lord 1980; Thissen & Mislevy, 2000) is an information-based method, and uses item response theory for module assignment. MFI is analogous to the maximum information item selection method in CAT, where a computer algorithm calculates the examinee's ability level based on the previously administered module(s), and then selects the module that maximizes information at his/her current ability estimate (Weissman et al., 2007). One can refer to Lord (1980) and/or Weissman et al. (2007) for more technical details.

Both two static methods define cut points (e.g., routing points or upper and lower bounds) for latent traits when routing examinees to the modules, but they differ in defining cut points. The first static rule of the module selection method is the approximate maximum information method (AMI) (Luecht, 2000) which is mainly used in criterion-referenced test administration (Zenisky, 2004). In the AMI routing method, a computer algorithm calculates the cumulative test information function (TIF) based on the previously administered modules (see the next section for discussion on information), and the TIFs of the current alternative modules (easy, medium, or hard). Then it adds each alternative module's TIF to the current cumulative TIF separately, and defines the intersection points of TIFs as the cut points (e.g., if the two intersection points of three TIFs are -1 and 1, the cut points are -1 and 1). Finally, the computer routes examinees to the alternative module that provides the highest information for the provisional latent trait of an examinee. For example, if $\theta \leq -1$, the examinee is routed to the easy module, if $-1 \leq \theta \leq 1$, the examinee is routed to the medium module, and if $\theta \geq 1$, the examinee is routed to the hard module. By nature, AMI requires an ability estimation method after each module, which brings additional complexity to the test administration. One can refer to Luecht (2000) and/or Zenisky (2004) for more technical details.

The second static rule of the module selection method is the defined population interval (DPI) or number-correct (NC), which is mainly used in norm-referenced test administration (Zenisky, 2004). This method is currently used in the revised version of GRE (Yan et al., 2014). The main goal in this method is to route a specific proportion of people to the modules to ensure an equal or nearly equal number (or proportion) of people draws each possible pathway (e.g., 33% of examinees routed to easy module, 34% of examinees routed to medium module, and 33% of examinees routed to hard module). For example, in a 1-3-3 panel structure, first the θ values corresponding to the 33rd and 67th percentiles of cumulative θ distribution are calculated (e.g., if the latent scores are normally distributed, θ-scores for the 33rd and 67th percentiles are -0.44 and 0.44, respectively). These are actually defined as the cut points in DPI method. Then these cut points are transformed to corresponding estimated true scores. Finally, examinees are routed to the one of the alternative modules based on the number of correct responses they got on the current module (e.g., people who got six or fewer items correct out of ten items are routed to the easy module, people who got seven or eight items correct are routed to the medium module, and people who got nine or ten items correct are routed to the hard module) (see Zenisky et al., 2010 for more details). Even if essentially or strictly parallel panels are built, it is always a good idea to check corresponding cut scores separately for each panel.

Compared to the MFI and AMI methods, the DPI is fairly straightforward to implement, but requires to an assumption on the distribution of theta scores to specify cut points prior to the test administration. Misrouting is more likely to occur when there is a huge discrepancy between the actual and assumed distributions. However, the number-correct method is a very understandable strategy by test takers (Hendrickson, 2007). Previous research has showed that in terms of routing decisions, DPI or NC is very practical and sufficient for a ca-MST design (Weissman et al., 2007). However, the information-based routing method is independent of theta distribution and may produce better outcomes than the static rules if there is discrepancy between the actual and assumed distributions. Yet, the choice of routing method is mainly determined by the purpose and consequences of the test (Zenisky et al., 2010).

Automated Test Assembly and Content Control

As explained, ca-MST is designed in such a way that test takers receive pre-constructed modules based on their performance on the previous module (Armstrong & Roussos, 2005). This means that each subsequent module has to match the current ability of a test taker. Thus, items must be carefully grouped in the modules. The most critical consideration here is to group items in modules based on the target information functions. Some have used the trial-error method and manually assembled modules (see Davis & Dodd, 2003; Reese & Schnipke, 1999), but it is quite difficult to satisfy all constraints (e.g., number of items, controlling content area) and to create parallel panels with manually assembled tests. Thus, it is always best to use a better strategy because automated test assembly (ATA) is a must in ca-MST (Luecht, 2000).

As Luecht describes, “The automated test assembly involves the use of mathematical optimization procedures to select items from an item bank for one or more test forms, subject to multiple constraints related to the content and other qualitative features” (2003, p.7). The optimum solution of ATA procedure ensures that items in modules and panels meet the desired constraints such as difficulty level, content control, word count, item and test overlap, and item format. As stated before, one has to decide ca-MST panel structure, total test length for a test taker, and total number of panels prior to ca-MST administration. Next, based on the ca-MST structure, the test designer has to determine the number of items in each module, and to pull a group of items from the given item bank that meet all desired constraints.

As shown by Luecht (1998), the automated test assembly provides a solution for maximizing the IRT information function at a fixed theta point. Let’s denote θ_0 as the fixed theta point and suppose we want a total of 24 items in the test. We first define a binary decision variable, x_i, (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;

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

where ξ_i represents the item parameters of item i (e.g., discrimination-*a*, difficulty-*b*, guessing-*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 from each content area. The automated test assembly is modeled to maximize

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

subject to

$$\sum_{i\epsilon C1}^{N} \quad x_i \geq 12 \tag{3}$$

$$\sum_{i\epsilon C2}^{N} \quad x_i \geq 12 \tag{4}$$

$$\sum_{i=1}^{N} \quad x_i \geq 24 \tag{5}$$

$$x_i \in (0,1),\ i = 1, \dots .N \tag{6}$$

which puts constraints on C_1, C_2, the total test length, and the range of decision variables, respectively. For illustration purposes, we provided an example of information functions for a 1-3 ca-MST panel design in Figure 5. As shown in this figure, the information functions for routing and the medium module peaks around θ=0, the information functions for the easy and hard modules peak around θ=-1 and θ=1, respectively.

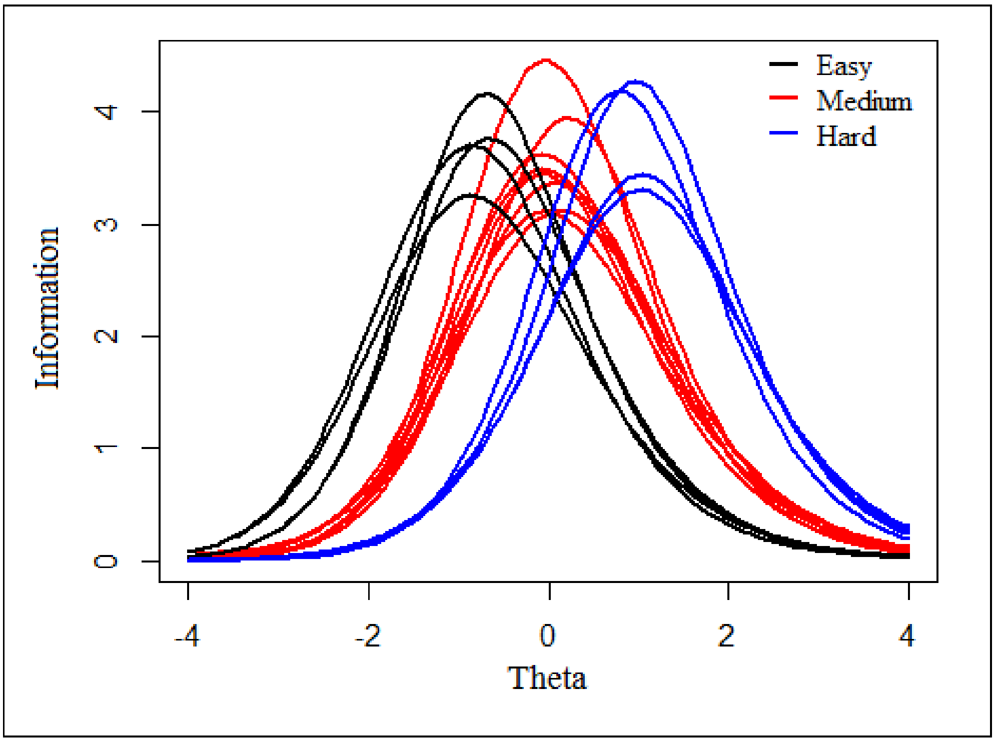

Figure 5. Test information functions across the modules at different difficulty levels

Adding content constraints in the equation is very important for meeting content requirements in the test. This is because for a successful and valid ca-MST, it is necessary to control content balancing so that each test taker takes has an equal and/or a similar number or percentage of items from each content area. It is well known that even if score precision is very high, as would be the case through proper adaptive algorithms in ca-MST, this does not necessarily ensure that the test has valid uses (Crocker & Algina, 1986). In other words, high score precision might not represent the intended construct if the content balance is not ensured. Furthermore, if all students are not tested on all aspects of the construct of interest, the test fairness is jeopardized.

For these reasons, any test form adapted to an examinee has to cover all related and required sub-content areas. Due to how it’s assembled, ca-MST allows greater control over test design and content. (i.e., ca-MST designers can determine item and content order within modules). 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, Chang, & Hau, 2003).

It is possible to add other constraints to the equations such as word count, especially if the test is a timed test. This is because students who receive too many items with high word counts will have a disadvantage even if the items are easy (Veldkamp & van der Linden, 2002).

One important issue when determining the structure of a ca-MST design, pulling items, and solving ATA problems is that it might not always be possible to extract a group of items from a large bank that are at the different difficulty levels. For example, if a researcher wants to build a 1-5 ca-MST panel design, and wants to allow more branching, she needs to have five modules in stage two (very easy, easy, medium, hard, and very hard). In such cases, the item bank must have groups of items with difficulty levels varying from very easy to very hard. In fact, the number of items needed for each module will increase for multiple panels. Otherwise, it will become impossible to find a proper solution for the desired structure. In short, the best solution of ATA for a desired ca-MST design can be found with a psychometrically rich item bank, because the quality of the item bank directly impacts especially complex ATA problems. It is also important to note that if item and/or test overlap is desired, a fewer number of items can be pulled from the item bank and placed in the modules. This can increase the likelihood of finding proper solution for an ATA problem.

There are some integer programming software programs used to solve ATA problems in ca-MST studies, such as IBM CPLEX (ILOG, Inc., 2006), CASTISEL (Luecht, 1998), LPSolve IDE[1], and IpSolve R package (Berkelaar, 2015). IBM Cplex is a commercial software, but a demo version allowing a limited number of constraints can be downloaded from www.ibm.com. The others are all non-commercial programs. Besides these, Microsoft Excel also provides a binary programming engine called Excel Solver Add-In. One can refer to Cor, Alves and Gierl (2009), and Diao and van der Linden (2011) for the informative works showing solving ATA problems in Excel and IpSolve R package, respectively.

After solving the ATA, the test designer has to place items into the modules and these modules into the panels. Two approaches are used to assign modules into the panels, bottom-up and top-down (Luecht & Nungester, 1998). In the bottom-up approach, all modules are built so as to meet module level specifications such as content and target difficulty. It is possible to think of each module as a mini test (Keng, 2008). Since modules constructed with this strategy are strictly parallel, the modules are exchangeable across the panels. In the bottom-down approach, the modules are built based on test level specifications. Since the modules are dependent and not parallel, they are not exchangeable across the panels.

Scoring and Ability Estimation Methods

The purpose of any test is to measure test takers' latent trait scores from responses given to a set of items (Crocker & Algina, 1986). At this point, ability estimation methods are used to calculate trait scores that represent student success. Even though classical test theory methods can be used for scoring, they are criticized by psychometric theoreticians and practitioners for producing test- and population-dependent outcomes and for focusing on true scores (Lord, 1980). Thus, we focus on IRT based scoring methods.

Like the routing method and automated test assembly, the ability estimation method is a key component in ca-MST administration. As discussed, if number-correct is the interested routing method, then ability estimation is not required when navigating examinees during the test. However, it is still required to estimate final ability level at the end of the test. If the ability cutoff-based or information-based routing methods are the interested routing methods, then it is necessary to use the

[1] http://web.mit.edu/lpsolve/doc/

ability estimation method both when navigating examinees and calculating final ability estimates at the end of the test.

There are two main ability estimation method groups that can be used to estimate both interim and final ability estimates: Bayesian methods and non-Bayesian methods. The most commonly used non-Bayesian method is maximum likelihood estimation (MLE) (Hambleton & Swaminathan, 1985). MLE begins with some a priori value (e.g., starting theta value) for the ability of the examinee and calculates the likelihood function with a given response pattern (e.g., 0101100111) for an ability level. The likelihood function estimates the probability of having that response vector and finds the value of theta that most likely results in that observed pattern. For example, if a person got many items wrong (e.g., response vector of 000010001), the likelihood function will tell us that this response pattern belongs to someone that has a very low latent trait. Then MLE finds the point that maximizes the likelihood of an examinee's item response pattern, and goes to the corresponding score on the latent trait scale. This score is the estimate of latent trait for a test taker. The first advantage of this method is that MLE is mathematically easier procedure compared to the other methods. Another advantage is that since item parameters are known in advance in ca-MST, estimations with MLE are unbiased compared to linear tests, in which parameters of items are unknown (Wang & Vispoel, 1998). One big disadvantage associated with MLE is that it does not provide an estimate for examinees that get all items right (i.e., perfect scores) or wrong (i.e., zero scores). This causes the likelihood function to infinitely increase and it becomes impossible to find the highest value on the likelihood function. This might be a serious problem for interim theta estimates in a ca-MST administration if there are few items in modules (Keller, 2000). One can refer to Hambleton and Swaminathan (1985) and/or Lord (1980) for technical details.

Two Bayesian methods that are commonly used are maximum a posteriori (Lord, 1986) and expected a posteriori (Hambleton & Swaminathan, 1985). Maximum a posteriori (MAP) is similar to MLE, but MAP specifies a prior distribution, multiples this prior distribution by the likelihood function, and then, does the same thing with the MLE. Most often, the prior distribution is chosen from a normal distribution. One big advantage of MAP is that it provides estimates for perfect and zero scores, outperforming MLE (Wang & Vispoel, 1998). One can refer to Hambleton and Swaminathan (1985) and/or Keller (2000) for technical details.

Another Bayesian method is expected a posteriori (EAP). EAP also specifies a prior distribution, but unlike the highest point on the likelihood function found in MAP, it finds the mean of the posterior distribution which represents the estimate of the latent trait. Unlike MLE and MAP, it is a non-iterative procedure, so it is easier to implement. Also, it does not assume normal prior distribution as MAP, and EAP outperforms MLE and MAP, meaning that it produces a lower standard error and bias than other ability estimation methods. One possible disadvantage with EAP is that if inappropriate prior distribution is specified, this affects the accuracy of outcomes. One can refer to Wainer and Thissen (1987) and/or Hambleton and Swaminathan (1985) for technical details.

One important issue for scoring and ability estimation is that when scoring the theta values, it is recommended to count all items an examinee received during the test (except pre-tested or seeded items) (Weissman et al., 2007). Then the estimated theta scores can be transformed to the desired rating scale (e.g., from 0 to 100). This is because the theta scores theoretically ranges from $-\infty$ to $+\infty$, but typically in practice ranges from -3 to 3 (Baker, 1992).

ON THE FLY COMPUTER ADAPTIVE MULTISTAGE TESTING

The main property of computer adaptive multistage testing is that selects the most appropriate module to an examinee and everyone works his/her own pace. As discussed before, the modules are pre-constructed "in house" and cannot change during the test. Due to this feature, the ca-MST aligns with linear tests and a test developer knows the prospective modules a test taker will receive during the test. However, in CAT, a test developer does not know which items a test taker will receive from a large item bank because CAT is a "on the fly" test. This feature of CAT has become an inspiration for ca-MST users, and on the fly multistage testing (OMST) has been proposed (Han & Guo, 2013).

Instead of administering pre-assembled modules in ca-MST, the OMST shapes item modules for theta interim values during the test. Due to this variation from ca-MST, some refer to on-the-fly MST as MST-shaping (MST-S), and ca-MST as MST-routing (MST-R) (see Han & Guo, 2014; Zheng & Chang, 2015). MST-S still holds other features of ca-MST such as item skipping and review, and it produces comparable results with ca-MST (Han &Guo, 2014). In summary, while ca-MST is somewhere between linear tests and CAT, MST-S is between ca-MST and CAT.

BOOKS AND SOFTWARE FOR COMPUTER ADAPTIVE MULTISTAGE TESTING

The first attempt to put together the work conducted on computer adaptive multistage testing was done by the Journal of Applied Measurement in Education, and a special volume (i.e., Volume 3) was organized in 2006. Then a special chapter (i.e., Chapter 18),"Multistage Testing: Issues, Designs, and Research," put together by April Zenisky, Ronald Hambleton, and Richard Luecht (2010) was published in the book of Elements of Adaptive Testing, edited by Wim J. van der Linden and Cees A. W. Glass (2010). Next, three ETS researchers, Duanli Yan, Alina A. von Davier, and Charles Lewis published the first ca-MST book, "Computerized Multistage Testing: Theory and Applications" (2014). To the best of our knowledge, this is the first and only book written for computer adaptive multistage testing.

The field of computer adaptive multistage testing is not very rich in terms of available computer programs or software yet. Two software programs can be used by researchers for their simulation studies: a) MSTGen (Han, 2013) and b) R (R Development Core Team, 2009-2016). The former, written by Kyung T. Han in 2013, is a Windows-based program, and is fairly user-friendly. It is available at no cost and can be downloaded from the author's website.[2] The current version of MSTGen supports both MST-R and MST-S. MSTGen supports three routing methods (maximum fisher information, matching-*b* value, and random module selection), as well as three theta scoring methods (MLE, MAP and EAP). MSTGen also supports creating and analyzing multiple panels. More information can be found in the user manual.[3]

The second program, R, is the most widely used open source program today and can be downloaded online.[4] R consists of user-created packages which include pre-written statistical commands. However, there is no special R package written for computer adaptive multistage testing yet. Thus, researchers have to write their own commands to run simulations for testing their ca-MST designs. Han and Kosinski (2014) provided an example R code that analyzes a ca-MST panel design found in Yan et. al. (2014) (see pages from 417 to 419 in Chapter 26). However, it is important to note that this code may not always serve researchers' intended study purposes. This is because this R code is limited to the maximum fisher information routing method, and is written for a special case where there must be equal number of items in all modules, which may not be always desired. This code also does not have a mechanism that prevents extreme jumps among the modules, which might cause aberrant response patterns.

Earlier this year, David Magis, the author of the computer adaptive testing R package called catR (Magis & Raiche, 2012) announced in the International Meeting of Psychometric Society Meeting in Asheville, N.C. that he is currently writing a ca-MST package named mstR. He plans to make it available for R users in fall 2017 (D. Magis, personal communication, July 11, 2016). MstR is going to support a variety of routing methods and ability scoring methods, and will not allow extreme jumps among the modules if user desires. The authors of this paper are also writing an R routine for

[2] http://www.umass.edu/remp/software/simcata/mstgen/

[3] http://www.umass.edu/remp/software/simcata/mstgen/MSTGen_Manual.pdf

[4] https://cran.r-project.org/mirrors.html

ca-MST analyses, and hope to release it in early 2017. The current version of the R routine can be requested from the first author of this paper.

DISCUSSION AND FUTURE RESEARCH ON CA-MST

Historically, student success has been mostly measured by linear test administration methods. However, these conventional methods are more accurate for measuring the ability of students at the medium ability level and less so for students at extreme ability points (Weiss & Kingsbury, 1984). They also produce longer tests (Segall, 2005). With modern advances in technology and measurement theory, adaptive testing has been proposed to cope with the problems posed by linear tests. So far, computerized adaptive testing has been the most popular and commonly preferred adaptive testing. During the last century, much research has been conducted on CAT. We strongly believe that even people outside of the educational and psychological measurement field are fairly familiar with CAT. Unfortunately, the same thing is not true for computer adaptive multistage testing.

This study does not argue against linear tests and/or computerized adaptive testing. Each test administration method has advantages and disadvantages when compared to the others. We believe that some superior features of linear tests, such as high test developer control on content balancing and item review, are vital. Similarly, some superior features of CAT, such as high measurement precision for theta estimates and shorter test lengths are also extremely important. To be honest, it is not possible to have all these features in the same test administration model.

This study argues that the computer adaptive multistage testing (ca-MST) “strikes a balance among adaptability, practicality, measurement accuracy, and control over test forms” (Zenisky, Hambleton, & Luetch, 2010, p.369) and combines all practical advantages of other test administration models. Despite these qualities, we believe that ca-MST has not been given enough attention and consideration, especially by researchers and practitioners in Turkey. We aim to arouse interest in ca-MST and encourage Turkish researchers to contribute this highly promising field.

Furthermore, we know that operational applications of adaptive testing are now being used in many European countries, the U.S., and Canada. Unfortunately, despite a large number of dedicated researchers and rapid advancements in technology, an operational example of adaptive testing has never been used in Turkey. Testing companies in the U.S. such as Educational Testing Service lead in adaptive testing. This was seen when the GRE switched from CAT to the ca-MST format, and the interest in ca-MST noticeably increased after 2011. The research reports released and distributed by ETS inspire researchers to find new research questions. We strongly believe that the Measuring, Selection and Placement Center (i.e. abbreviated as OSYM in Turkish) can have the same impact on the researchers in Turkey by releasing an operational example of adaptive testing and distributing research reports associated with it.

Ca-MST has fewer number of routing methods. Whereas some item selection methods used in CAT such as Kullback-Leibler (Chang & Ying, 1996), maximum likelihood weighted information (Veerkamp & Berger, 1997), the maximum posterior weighted information (van der Linden, 1998) can be easily adopted and modified, and then used as a routing method in the ca-MST environment. Furthermore, some special and popular topics such as differential item functioning, item parameter drift, item copying, and cognitive diagnostic models should be investigated in the ca-MST environment. Also, a computer software comparison study across MSTGen and R should be conducted to compare the effectiveness, usefulness, and accuracy of these two software programs. And last but not least, another comparison study looking at IBM Cplex, CASTISEL, LPSolve, and IpSolve R package should be conducted to explore the effect of integer programming software when creating ca-MST panel designs. Researchers can always contact the authors with questions and for help regarding ca-MST design, software, coding, ATA, etc.

REFERENCES

Angoff, W. H., & Huddleston, E. M. (1958). *The multi-level experiment: a study of a two-level test system for the College Board Scholastic Aptitude Test* (SR-58-21). Princeton, New Jersey: Educational Testing Service.

Armstrong, R. D. & Roussos, L. (2005). *A method to determine targets for multi-stage adaptive tests*. (Research Report 02-07). Newtown, PA: Law School Admissions Council.

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

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.

Becker, K. A., & Bergstrom, B. A. (2013). Test administration models. *Practical Assessment, Research & Evaluation*, *18*(14), 7.

Belov, D. I., & Armstrong, R. D. (2005). Monte Carlo test assembly for item pool analysis and extension. *Applied Psychological Measurement*, *29*, 239-261.

Berkelaar, M. (2015). Package ‘lpSolve’.

Bridgeman, B. (2012). A Simple Answer to a Simple Question on Changing Answers. *Journal of Educational Measurement*, *49*, 467-468.

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.

Cor, K., Alves, C., & Gierl, M. (2009). Three Applications of Automated Test Assembly within a User-Friendly Modeling Environment. *Practical Assessment, Research & Evaluation*, *14*(14).

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

Cronbach, L. J., & Glaser, G. C. (1965). Psychological tests and personnel decisions. Urbana, IL: University of Illinois Press.

Crotts, K. M., Zenisky, A. L., & Sireci, S. G. (2012, April). *Estimating measurement precision in reduced-length multistage-adaptive testing.* Paper presented at the meeting of the National Council on Measurement in Education, Vancouver, BC, Canada.

Davey, T., & Y.H. Lee. (2011). Potential impact of context effects on the scoring and equating of the multistage GRE Revised General Test. (GRE Board Research Report 08-01). Princeton, NJ: Educational Testing Service.

Davis, L. L., & Dodd, B. G. (2003). Item Exposure Constraints for Testlets in the Verbal Reasoning Section of the MCAT. *Applied Psychological Measurement*, *27*, 335-356.

Diao, Q., & van der Linden, W. J. (2011). Automated test assembly using lp_solve version 5.5 in R. *Applied Psychological Measurement*, DOI: 0146621610392211.

Dubois, P. H. (1970). *A history of psychological testing*. Boston: Allyn & Bacon

Eignor, D. R., Stocking, M. L., Way, W. D., & Steffen, M. (1993). Case studies in computer adaptive test design through simulation. *ETS Research Report Series*, *1993*(2).

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications* (Vol. 7). Springer Science & Business Media.

Han, K. T. (2013). " MSTGen": Simulated Data Generator for Multistage Testing. *Applied Psychological Measurement*, *37*, 666-668.

Han, K. T., & Kosinski, M. (2014). Software Tools for Multistage Testing Simulations. In *Computerized Multistage Testing: Theory and Applications* (pp. 411-420). Chapman and Hall/CRC.

Han, K.T., & Guo, F. (2013). *An Approach to Assembling Optimal Multistage Testing Modules on the Fly* (Report No. RR-13-01). Reston, Virginia: Graduate Management Admission Council. Retrieved from GMAC website: http://www.gmac.com/market-intelligence-and-research/research-library/validity-and-testing/research-reports-validity-related/module-assembly-on-the-fly.aspx

Hendrickson, A. (2007). An NCME instructional module on multistage testing. *Educational Measurement: Issues and Practice*, *26*(2), 44-52.

ILOG. (2006). ILOG CPLEX 10.0 [User’s manual]. Paris, France: ILOG SA.

Keller, L. A. (2000). Ability estimation procedures in computerized adaptive testing. *USA: American Institute of Certified Public Accountants-AICPA Research Concortium-Examination Teams*.

Keng, L. & Dodd, B.G. (2009, April). *A comparison of the performance of testlet based computer adaptive tests and multistage tests*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Keng, L. (2008). *A comparison of the performance of testlet-based computer adaptive tests and multistage tests* (Order No. 3315089).

Kim, H., & Plake, B. S. (1993, April). *Monte Carlo simulation comparison of two-stage testing and computerized adaptive testing*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Atlanta, GA.

Leung, C.K., Chang, H.H., & Hau, K.T. (2002). Item selection in computerized adaptive testing: Improving the a-Stratified design with the Sympson-Hetter algorithm. *Applied Psychological Measurement*, *26*(4), 376–392.

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).

Lewis, C., & Sheehan, K. (1990). Using Bayesian decision theory to design a computerized mastery test. *Applied Psychological Measurement*, *14*(4), 367-386.

Lord, F. M. (1971). A theoretical study of two-stage testing. *Psychometrika*, 36, 227–242.

Lord, F. M. (1974). Practical methods for redesigning a homogeneous test, also for designing a multilevel test. Educational Testing Service RB-74–30.

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

Lord, F. M. (1986). Maximum likelihood and Bayesian parameter estimation in item response theory. *Journal of Educational Measuremen*t, *23*, 157–162.

Luecht, R. M. & Sireci, S. G. (2011). A review of models for computer-based testing. Research (Report No: 2011-12). New York: The College Board. Retrieved from website: http://research.collegeboard.org/publications/content/2012/05/review-models-computer-based-testing

Luecht, R. M. (1998). Computer-assisted test assembly using optimization heuristics. *Applied Psychological Measurement*, *22*, 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. (2003, April). *Exposure control using adaptive multi-stage item bundles*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL.

Luecht, R. M., & Nungester, R. J. (1998). Some Practical Examples of Computer- Adaptive Sequential Testing. *Journal of Educational Measurement*,*35*, 229-249.

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

Luecht, R. M., Nungester, R.J., & 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.

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

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

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.

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

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 *Elements of adaptive testing* (pp. 151-165). Springer New York.

Schnipke, D. L., & Reese, L. M. (1999). A Comparison [of] Testlet-Based Test Designs for Computerized Adaptive Testing. Law School Admission Council Computerized Testing Report. LSAC Research Report Series.

Segall, D. O. (2005). Computerized adaptive testing. *Encyclopedia of social measurement*, *1*, 429-438.

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.

Thompson, N. A., & Weiss, D. J. (2011). A framework for the development of computerized adaptive tests. *Practical Assessment, Research & Evaluation*,*16*(1), 1-9.

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

van Der Linden, W. J., & Chang, H. H. (2003). Implementing content constraints in alpha-stratified adaptive testing using a shadow test approach. *Applied Psychological Measurement*, *27*, 107-120.

van der Linden, W.J., Jeon, M., & Ferrara, S. (2011). A paradox in the study of the benefits of test-item review. *Journal of Educational Measurement*, *48*, 380-398.

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

Veldkamp, B. P., & van der Linden, W. J. (2002). Multidimensional adaptive testing with constraints on test content. *Psychometrika*, *67*, 575-588.

Wainer, H., & Thissen, D. (1987). Estimating ability with the wrong model. *Journal of Educational and Behavioral Statistics*, *12*, 339-368.

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., & Mislevy, R. J. (2000). *Computerized adaptive testing: A primer*. Routledge.

Wang, T., & Vispoel, W. P. (1998). Properties of ability estimation methods in computerized adaptive testing. *Journal of Educational Measurement*, *35*, 109-135.

Wang, X., Fluegge, L., & Luecht, R.M. (2012, April). *A large-scale comparative study of the accuracy and efficiency of ca-MST panel design configurations*. Paper presented at the meeting of the National Council on Measurement in Education, Vancouver, BC, Canada.

Weiss, D. J. (1973). *The stratified adaptive computerized ability test* (Research Report 73-3). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

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

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

Yan, D., von Davier, A. A., & Lewis, C. (Eds.). (2014). *Computerized multistage testing: Theory and applications*. 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: Springer.

Zheng, Y., & Chang, H. H. (2015). On-the-fly assembled multistage adaptive testing. *Applied Psychological Measurement*, *39*(2), 104-118.

Zheng, Y., Nozawa, Y., Gao, X., & Chang, H. H. (2012). *Multistage Adaptive Testing for a Large-Scale Classification Test: Design, Heuristic Assembly, and Comparison with Other Testing Modes. ACT Research Report Series, 2012 (6).* ACT.

GENİŞ ÖZET

Öğrenci yeteneğini veya başarısını ölçmek birçok testin öncelikli amacıdır. Bu amaç doğrultusunda en çok başvurulan yöntem klasik yöntem diye adlandırılan kağıt-kalem formatındaki testlerdir. Bu yöntemin en büyük avantajlarından biri test formlarının hazırlanmasında testi organize eden kişiye test içeriğini oluşturmada büyük kolaylık sağlamısıdır. Bunun nedeni testi hazırlayan kişinin test içeriğini, testteki maddelerin sırasını, soru sayısını istediği gibi belirleyebilmesidir. Ayrıca bu yöntem soru bankası oluşturmayı gerektirmediği için daha az maliyetlidir. Ancak bu yöntem öğrenci başarısını ölçmede yüksek yanlılık (hata) ürettiği ve test uzunluğu nediyle oldukça eleştirilmektedir. Bu sebeble bilgisayar ortamında uygulanan bireye uyarlanmış testler geliştirilmistir.

Günümüzde en yaygın kullanılan ve bilinen bireye uyarlanmış test; madde bazında bireye uyarlanmış testlerdir. Bu test yönteminin genel çalışma prensibi şu şekildedir; bilgisayar bireye başlangıç için bir soru verir (genellikle orta zorluk derecesinde), bireyin bu maddeye verdiği cevap sonrasında, bilgisayar bireyin yetenek seviyesini hesaplar, hesaplanan yeni yetenek seviyesine göre başka bir soru verir ve yetenek seviyesini günceller. Bu işlem test sonuna kadar devam eder. Testi sona erdiren mekanizma testi organize eden kişi tarafından belirlenir ve a) süre, b) önceden belirlenen standart hata (örneğin kişinin yetenek seviyesi 0.3 hata oranıyla hesaplandığında testi durdur) veya c) önceden belirlenen test uzunluğu olabilir (örneğin her bir birey 30. sorusunu aldığında testi durdur). Madde bazında bireye uyarlanan testin en büyük avantaji yetenek seviyesini minimum hata ile hesaplayabilmesi ve soru sayısını %50 oranında azaltmasıdır. Kısacası daha az soru ile daha iyi bir ölçmeyi gerçekleştirmesidir. Ancak bu yöntemin de kendine özgü dezavantajları bulunmaktadır. Belki de en ciddi olanı yüksek maliyet gerektirmesidir çünkü bu test yöntemi soru bankası oluşturmayı zorunlu kılmaktadır. Bu yöntemin diğer dezavantaji bireylere önceki sorulara gerip dönüp cevaplarını değiştirme veya gözden geçirme şansı tanımaması ve her bir sorunun cevaplanmasını zorunlu kılmasıdır.

Bilgisayar ortamında uygulanan diğer metod ise bireye uyarlanmış çok aşamalı testlerdir. Bireye uyarlanmış çok aşamalı testler isminden anlaşılacağı üzere çesitli bölümlerden oluşur ve her bir bölümde modül adı verilen farklı zorluk derecesinde soru kümeleri bulunmaktadır. Bu bağlamda her bir modülü mini bir test olarak düşünmek mümkündür. İlk aşamada genellikle tek modül bulunur ve yönlendirme modülü (routing module) olarak adlandırılır. Fakat diğer aşamalarda farklı zorluk derecelerinde birden fazla modül bulunmaktadır. Bireye uyarlanmış çok aşamalı testlerin çalışma prensibi şu şekildedir; birey önce yönlendirme modülünü alır ve bu modülde göstermiş olduğu performasa bağlı olarak ikinci aşamada düşük, orta veya yüksek zorluk seviyesindeki modülü alır ve ikinci aşamayı tamamlar. Bu işlem birey tüm aşamaları bitirinceye kadar devam eder. Anlaşıldığı üzere madde düzeyindeki bireyselleşmeden ziyade, bireyselleşme modül düzeyinde gerçekleşmektedir. Testi hazırlayan kişi testteki aşama sayısını kendisi belirleyebilir ve her aşamaya istediği kadar modüle yerleştirebilir. Şekil 2 ve 3 çok aşamalı testlerin yapısını gösteren birer örnektir. Bu şekillerde sadece bir panel gösterilmektedir. Halbuki çok aşamalı testler birbirine paralel olan çok sayıdaki panelden oluşmaktadır ve bireyler herhangi bir panele rastgele atanır. Bu panellerin bireylere maruz bırakılma veya kullanım oranını belirli seviyede tutma ve test güvenliğini sağlama açısından oldukça önemlidir. Bireye uyarlanan çok aşamalı testlerin madde düzeyinde bireye uyarlanan testlere göre en büyük avantajlarından biri bireylere her bir modül içerisindeki önceki sorulara geri dönüp cevapları gözden geçirmesine imkan vermesidir. Ancak bireylerin bir önceki aşamadaki modülde aldıkları sorulara geri dönmesine izin verilmez. Diğer bir avantajı ise testi hazırlayan kişiye test içeriği üzerinde daha fazla imkan vermesidir. Bunun nedeni modül içerisindeki sorular testi hazırlayan tarafından önceden belirlenir, istenilen sayıda ve sırada soru yerleştirilebilir. Bireye uyarlanan çok aşamalı testler bireylerin yetenek seviyelerini ölçmedeki hata derecesi açısından madde bazında bireye uyarlanan testlerden kötü, kağıt kalem formundaki testlerden iyidir. Fakat madde bazında uyarlamanın olduğu testlerde karşılaşılan yüksek maliyet gibi sorunlar çok aşamalı testlerde de görülmektedir. Bu nedenlerden dolayı bireye uyarlanan çok aşamalı testlerin diğer iki testin avantajlarını ve dezavantajlarını içermektedir. Çok aşamalı bir test yönlendirme metodu denilen adaptasyonu gerçekleştiren mekanizmayi kullanmayı gerektirir. Yönlendirme metodu bireylerin bir sonraki aşamada hangi modülü alacağını belirler ve bireylerin yetenek seviyelerinin hesaplanmasında çok önemli rol oynar. Bu nedenle çok aşamalı testlerin en önemli unsurlarından biridir. Çalışma en yaygın kullanılan üç farklı yönlendirme metodunu detaylı şekilde açıklamaktadır. Bunlar doğru sayısına göre, hesaplanan yetenek seviyesine göre ve test bilgisine (test information) göre yönlendirme metodlaridir. Çok aşamalı testlerdeki önemli noktalardan biri de “automated test assembly” (ATA) olarak adlandırılan madde havuzundaki modülleri oluşturmaya yarayan test toplama/bir araya getirme metodudur. ATA farklı zorluk derecelerindeki maddeleri bir araya getirmeye yarar. Bunu yaparken farklı her bir modül içerisinde o teste ait farklı alt konulara (denklemler, fonksiyonlar, toplama/çıkarma vs.) ait soruları bir araya getirmeye dikkat edilmelidir. Çalışmada panellerin ve modüllerin nasıl oluşturulduğunu detaylıca anlatmaktadır. Diğer bir önemli unsur ise yetenek seviyesini hesaplama metodudur. En yaygın kullanılan metodlar “en çok olabilirlik (MLE), sonsal maksimum kestirim (MAP) ve sonsal beklenti kestirimi (EAP) metodlarıdır. Her bir metodun sınırlılıkları ve üstün yönleri çalışmada detaylıca anlatılmıştır. Bireye uyarlanan çok aşamalı testler üzerine yazılmış kitaplar ve tasarlanmış bilgisayar yazılımları çalışmada tanıtılmıştır. Çalışmada ayrıca geçmişten günümüze yapılan çalışmalar özetlenmiş ve gelecekte yapılabilecek çalışmalar tartışılmıstır.

Madde bazında bireye uyarlanan testlerin ilk adımları 1905 yılında atılmış olup yüzyılı aşkın bir geçmisi bulunmaktadır. Günümüze gelene kadar üzerinde çokça çalışma yapılmış ve ciddi ilerleme sağlanmıştır. Bu yüzden eğitimde ve psikolojide ölçme alanı dışındaki araştırmacılar tarafından bile birçok yönü itibariyle bilinmektedir. Fakat modül bazında bireye uyarlanan veya bireye uyarlanan çok aşamalı testler diğer test yöntemlerine göre çok daha yeni olduğu için birçok araştırmacı tarafından bilinmemekte veya kısmen bilinmektedir. Bu çalışmanın amacı bireye uyarlanan çok aşamalı testleri tüm yönleriyle incelemek, temel prensiplerini anlatmak, gelinen noktayı ve yapılan çalışmaları araştırmacılar için özetlemektir. Türkiye'de bu alanda yapılan çalışmalar çok kısıtlı olduğu icin hedefimiz Türkiye'deki araştırmacıların dikkatini bu alana çekmek ve katkıda bulunmalarına teşvik etmektir. Kuşkusuz bireye uyarlanan çok aşamalı testler avantajlari itibariyle bireye uyarlanan madde bazındaki testlerin yerini almaya aday. Bu nedenle de cok aşamalı testlere olan ilgi gelecekte daha da artacaktır. Özellikle 2011 yılında Graduate Record Examination (GRE) madde bazında bireyselleştirmeden modül bazında bireyselleştirmeye geçtikten sonra, çok aşamalı testlere olan ilgi gözle görünür şekilde artmıştır. Bunu yayınlanan akademik araştırma ve yapılan simülasyon çalışmalarının sayısından rahatça görmekteyiz. Benzer ilginin Türkiye'de de artması için yapılabilecek çalışmalara değinilmiştir.