

Analyzing Different Module Characteristics in Computer Adaptive Multistage Testing

Melek Gulsah Sahin ^{1,*}

¹Gazi University, Gazi Education Faculty, Department of Educational Sciences, Turkey

ARTICLE HISTORY

Received: Jan 19, 2020

Revised: Mar 6, 2020

Accepted: Apr 19 2020


KEYWORDS

ca-MST,
Panel design,
Module length,
Item discrimination
sequence,
Difficulty difference
condition

Abstract: Computer Adaptive Multistage Testing (ca-MST), which take the advantage of computer technology and adaptive test form, are widely used, and are now a popular issue of assessment and evaluation. This study aims at analyzing the effect of different panel designs, module lengths, and different sequence of a parameter value across stages and change in b parameter range on measurement precision in ca-MST implementations. The study has been carried out as a simulation. MSTGen simulation software tool was used for that purpose. 5000 simulees derived from normal distribution ($N(0,1)$) were simulated. 60 different conditions (two panel designs (1-3-3; 1-2-2), three module lengths (10-15-20), 5 different a parameter sequences (“0.8; 0.8; 0.8” - “1.4; 0.8; 0.8”-“0.8;1.4; 0.8” - “0.8; 0.8;1.4” - “1.4; 1.4; 1.4”) and two b parameter difference (small; large) conditions) were taken into consideration during analysis. Correlation, RMSE and AAD values of conditions were calculated. Conditional RMSE values corresponding to each ability level are given in a graph. Dissimilar to other studies in the literature, this study examines b parameter difference condition in three-stage tests and its interaction with a parameter sequence. Study results show that measurement precision increases as the number and length of the modules increase. Errors in measurement decrease as item discrimination values increase in all stages. Including items with a high value of item discrimination in the second or last stage contributes to measurement precision. In extreme ability levels, large difficulty difference condition produces lower error values when compared to small difficulty difference condition.

1. INTRODUCTION

In line with the fact that computer technology has led to various differences in all domains of life, it has changed the way cognitive/affective tests are carried out. Traditional paper-and-pencil tests have gradually been replaced by computer based testing (CBT) in time. When the qualities of tests that are administered on the basis of CBT are considered, it can clearly be observed that there is a version which prescribes all individuals to take the same form as well as the other version which assigns the use of an adaptive form (computerized adaptive testing-CAT) that makes it possible to determine the test items in accordance with the abilities of individuals who take the test. In the background, the adaptive form performs some operations

CONTACT: Melek Glah ahin ✉ mgulsahsahin@gazi.edu.tr  Gazi University, Gazi Education Faculty, Department of Educational Sciences, Turkey

ISSN-e: 2148-7456 /© IJATE 2020

that are based on Item Response Theory (IRT), which is a testing theory that increases the measurement precision about the estimation of item and ability parameters and that brings about a great number of advantages in terms of implementation. Computer Adaptive Testing (CAT) has been frequently preferred for use thanks to the numerous advantages it offers and it has been discussed in many studies up until now. CAT has item level adaptation as its basis, because the algorithms that are created for item selection build up each test form while the test is being taken by the examinee via controlling an item, estimating a tentative score, and later choosing the next item to be used from the active item bank through making use of some specific statistical optimization criteria (Luecht & Nungester, 1998). As a result, examinees are not allowed to review their responses to previous items when a CAT implementation is adopted. Furthermore, the item exposure rates of some items might be high although the item sets that examinees come across differ from each other. Unfolding plenty of items, no matter how many examinees see them, can influence the accuracy and validity of test scores if the examinees that will take the test in the future have an opportunity to see the test items before testing (Rotou, Patsula, Steffen, & Rizavi, 2007). Another limitation that is caused by the fact that examinees take different test items is that it turns out to be impossible to examine each test form with quality assurance purposes before testing (Luecht & Nungester, 1998).

Computer Adaptive Multi-stage Testing (ca-MST) has not only taken the advantages of computer technology as well as adaptive forms but also found a way to overcome the problems of delivering a different set of items to each examinee. The ca-MST has been widely implemented thanks to these qualities, and it is one of the issues that are frequently studied in the field of measurement and evaluation nowadays. Because of this reason, there are some assessments that have replaced CAT versions with ca-MST versions, such as the National Assessment of Educational Progress (NAEP) and the Graduate Record Examinations (GRE) (Zeng, 2016; Zheng, Nozawa, Gao, & Chung, 2012). The major distinction between ca-MST versions and CAT versions is that ca-MST prescribes examinees to take a set of pre-constructed sub-test which matches their tentative ability estimates all the time (Hendrickson, 2007). However, when a CAT is used, only a single item is selected to match the ability estimates (Zeng, 2016). These pre-assembled sub-tests are called modules and ca-MST make use of these fundamental building blocks (Leucht & Sireci, 2011). In brief, it can be stated that the main difference between ca-MST and CAT is that ca-MST is a module adaptive test, not an item level adaptive test. The literature review shows that CAT and ca-MST are frequently compared against the backdrop of some certain qualities. As a consequence, ca-MST stays between linear test forms (paper and pencil testing and computer-based testing) and conventional item-level CAT (Hendrickson, 2007; Leucht & Nungester, 1998; Sadeghi & Khonbi, 2017; Sarı, Sarı, & Huggins Manley, 2016).

According to Leucht and Sireci (2011), ca-MST has some other advantages besides paying regard to content specifications and item exposure issues. These advantages include, but are not limited to, enabling examinees to review test items included in the same test, simplifying the test format, obtaining test results close to CAT versions especially when long tests and different contents are in question, simplifying the expensive programs that are used for test development and administration, being able to fix “information structure” that is necessary for each panel and reproducing it among panels, making it possible to examine the quality of panels before the test is administered to the test-takers (Hadadi & Leucht, 1998; Leucht & Nungester, 1998; Leucht, 2000; van der Linden, 2005; Patsula, 1999; Schnipke & Reese, 1999; Zenisky & Jodoin, 1999; Zenisky & Hambleton, 2014).

1.1. ca-MST Components

There are four basic test design/administration concepts in ca-MST: (1) modules, (2) panels, (3) stages, and (4) pathways (Leucht, 2000). Modules are units that are homogeneous in terms

of item difficulty. Each module can be structured in line with a specific content and statistical characteristics before examinees take the test (Leucht & Nungester, 1998). The length of the modules is based on the nature of the test, so a module can change between a small size (five to ten items) and large size (50 to 100 items). They can also differ in length according to stages and average difficulty (Leucht, 2000). There is a certain statistical target that should be met by each series of modules in terms of a psychometric perspective; this target can be described as a prescribed level of measurement precision within a specific region of the score scale (i.e., an IRT test information target) (Leucht & Sireci, 2011).

Test assembly requires a process where item modules are grouped so as to form test administration units that are called "panels" in accordance with stage and difficulty level (Leucht & Nungester, 1998; Leucht & Sireci, 2011). A panel is a specific combination of the modules that have to meet the pre-supposed requirements of content and other qualitative test features besides other explicit statistical targets (Leucht & Nungester, 1998). There is a natural hierarchical arrangement which designates panels which own multiple modules and modules that own multiple items (Leucht & Sireci, 2011). The ca-MST panels are divided into two or more stages and each module included in the panel is assigned to a specific stage. Two- or three-stage designs are widely used (Park, 2015). A ca-MST stage may have more than one module, and each module may address a different proficiency level (e.g., one easy, one moderate, and one hard module, each of which is aimed at a particular range of examinee abilities) (Leucht & 1998). A panel configuration refers to a simple sequence of integers that indicate the number of stages and amount of adaptation that are possible in a specific panel design (Leucht & Sireci, 2011). A ca-MST design that consists of more than two stages puts forth that how an examinee performs on the second stage of the test helps routing the examinee to a third stage module later on.

A simple panel structure can be seen in [Figure 1](#) below. This is called a 1-3-3 panel design. In this design, modules with different difficulty levels are created and specified in the figure. The first stage includes a module with a difficulty level of medium (M) or average. On the other hand, the item difficulty levels of modules in the second and third stages range from easy (E) to hard (H). Here, there are seven possible pathways designated for examinees: M1+E2+E3, M1+E2+M3, M1+M2+E3, M1+M2+M3, M1+M2+D3, M1+H2+M3, and M1+H2+H3 (Leucht & Sireci, 2011).

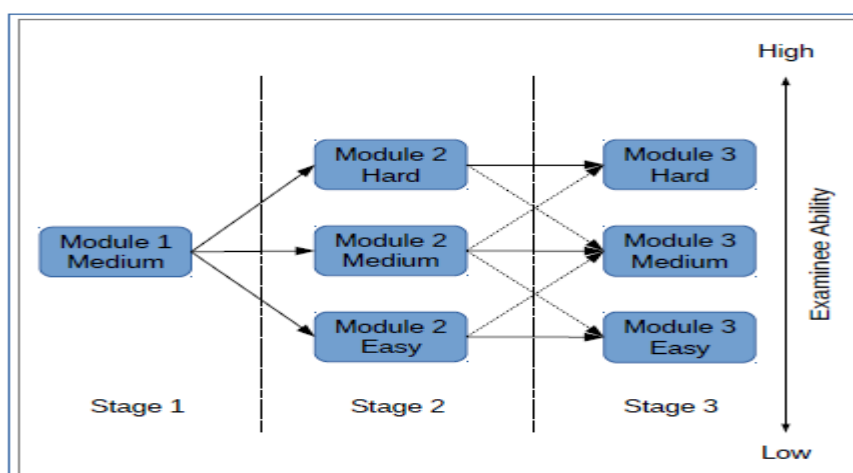


Figure 1. Structure of 1-3-3 panel design of ca-MST

Only one module from each stage is administered for each examinee during an actual test delivery process. Examinees having different abilities are routed to different modules. For

example, an examinee whose ability estimation is at a high level may be assigned a module that includes more difficult items or negatively skewed test (module) information functions (Zenisky & Jodoin, 1999). In [Figure 1](#), Medium-1, Hard-2 and Hard-3 modules are created for a high proficiency examinee.

Principally, some variables should be considered for a ca-MST design. Zenisky (2004) classifies the variables that are necessary for a ca-MST design as follows: i) basic structure variables such as total number of items and total number of stages included in the test, ii) variables of test and module assembly such as difficulty of the first-stage module, the number of the relative difficulty of the modules in the following stage, and content balance and other limitations; and iii) variables related to administration which influence the efficacy and implementation of ca-MST, including strategies of routing and ability estimation methods. In addition to the variables listed above, some other variables that are considered in ca-MST implementations include panel design considerations such as the quality of the item pool, distribution of the difficulty and item discrimination for each content, and the number of modules included in each stage (Han & Guo, 2013; Leucht, Brumfield, & Breithaupt, 2006; Park, 2015; Xing & Hambleton, 2004).

1.2. Aim of the Study

In this study, basic structure as well as test and module assembly variables such as different panel designs, change in b parameter difficulty level of modules, different distributions of item discrimination in the stages and test length have been examined. The aim of this study is to analyze the influence of the specified conditions on measurement precisions. As it is seen in the literature, there are studies which have focused on the effect of some variables such as the test length, dichotomous items and polytomously-scored items on ca-MST test performance (Jodoin, Zenisky, & Hambleton, 2006; Leucht & Nungester, 1998; Patsula, 1999; Xing & Hambleton, 2004). Patsula (1999) carried out a study that examined panel designs including different module numbers and underlined the fact that it is important to examine the number of modules included in each stage instead of the length of modules. Another variable that influences measurement precision is test information function (TIF) which is a degree of measurement precision demanded in the various regions of the ability scale included in the test to be administered. TIF also plays an important role in ensuring the consistency of results obtained from ca-MST against time and across panels (Leucht, 2000; Leucht & Nungester, 1998).

While TIF values are sometimes expected to be maximum at specified decision points in accordance with the aim of the test (i.e., when the test aims at classification), they are expected to be flat in between specific intervals (i.e., a test needs to assess ability across a wide range of theta scale) (Verschoor & Eggen, 2014; Park, 2015). As each module is constructed before the test is implemented in ca-MST design, it is not possible to control TIF values later. This case is directly related to the reliability of the test results. If the number of test items included in the routing module is low, there might be some items that can be answered correctly by guessing. In this case, it turns out to be more important to make ability estimations regarding examinees' performances with fewer mistakes in the following stages. In such cases, if the modules included in the following stages which examinees are routed focus on a narrow region of abilities, this may cause misrouting (Kim & Moses, 2014). This is especially very important in two-stage designs as there is a single adaptation point. In tests that consist of different numbers of stage (three stages or more), it is important to examine the effect of the intervals of TIF values specified for each stage on measurement precision as well as test assembly. In the literature, there are studies in which modules are constructed considering different TIF distributions or difficulty parameter values and the results are compared accordingly (Kim, Chung, Dodd, & Park, 2012; Kim & Moses, 2014; Kim, Moses, & Yoo, 2015).

Another essential point in constructing the modules is paying attention to item discrimination. There is no doubt that there is a relationship between item discrimination and test reliability. In the literature, there are studies which examine the effect of item discrimination on creating item pool and routing module (Boztunç Öztürk, 2019; Xing & Hambleton, 2004).

1.3. Significance of the Study

It is recommended in the literature to examine the change in ranges of difficulty levels related to b parameters for modules in ca-MST implementations (Leucht, Brumfield, & Breithaupt, 2016) and to deal with this change together with the impact of different levels of item discrimination (Kim & Moses, 2014; Kim et al., 2015). During the literature review, the researcher has not come across any study that examines the change in item discrimination of modules included in different stages. Therefore, it is thought that it will be worthwhile to examine the interaction of modules that are constructed depending on small and large b parameter (difficulty) level differences with different values of average discrimination within the framework of this study. Another significance of this study is that it will take the advantage of three-stage panel design (1-3-3 and 1-2-2) unlike other studies in the literature (Kim & Moses, 2014; Kim et al., 2015). The reason why these panel designs are chosen is that 1-2-2 ca-MST structure is popular for classification testing, while 1-3-3 design is the most commonly preferred research for ability estimation testing (Jodoin et al., 2006; Park, 2015; Zenisky, 2004). Furthermore, the researcher aims at contributing to the literature by examining the module length together with module difficulty and module discrimination values. Also, some suggestions will be provided for test operators to use in practice in light of the findings that will be obtained at the end of this study

2. METHOD

2.1. ca-MST Panel Assembly

This study examines the panel design of 1-3-3, which is the most frequently preferred one in the literature (Chen, 2010; Hambleton & Xing, 2006; Jodoin et al., 2006; Leucht & Nungester, 1998; Leucht et al., 2006; Park, 2015; Patsula, 1999; Zenisky, 2004). Patsula (1999) has stated that the change in the number of modules, not stages, produces a difference in terms of measurement precision. This study also addresses 1-2-2 panel design, which is also three-stage but has a different number of modules (Chen, 2010; Patsula, 1999; Zenisky, 2004). The second variable that is considered within the framework of this study is module length assignment. It is preferred to have a condition where each stage includes equal numbers of items. Chen (2010) has underlined that ca-MST studies generally make use of items ranging from 33 to 60 in number. Within the framework of this study, the module length is chosen to be 10, 15 and 20 items, whereas test length is decided to be 30, 45 and 60 items in total for the purpose of observing change in tests that have an average length on one side and long tests on the other side.

The third variable that is varied in this study is item discrimination. The study aims at designating at which stages the average discrimination values of items that are included in the modules can be high or low in a three-stage ca-MST implementation. It is seen that item discrimination has an average value of 0.75-0.85 and SD value of 0.27-0.30 in studies that have been carried out with parameters obtained from a real pool (Kim & Moses, 2014; Kim et al., 2015, Patsula, 1999; Zheng & Chang, 2015). Hambleton and Xing (2004) carried out a study in which they identified item quality as poor ($\bar{x}=0.60$), average ($\bar{x}=1.00$) and best corresponding to average ($\bar{x}=1.40$) according to a parameter value. In this study, a parameter was addressed as average ($\bar{x}=0.80$; $SD=0.25$) and high ($\bar{x}=1.40$; $SD=0.25$) for each stage. Within the framework of this study, a parameter average of items included in the modules in each stage for a three-stage model are addressed with five different conditions which can be

listed respectively as average-average-average, high-average-average, average-high-average, average-average-high and high-high-high.

Small b parameter difficulty level and large b parameter difficulty level were selected while constructing the modules included in stages for both panel designs. Literature review shows that there are studies which are conducted with two-stage tests and examine difficulty difference conditions. Kim, Moses and Yoo (2015) carried out a simulation study in which they designated the theta values as ($b = -0.5$, $b = 0.0$ and $b = 0.5$) in small difficulty difference condition and as ($b = -0.5$, $b = 0.0$ and $b = 0.5$) in large difficulty difference conditions for the second stage. They specified the difficulty difference between easy modules as 0.5 in small and large conditions. The same design was used for the difficulty difference between difficult modules. In addition, medium module was set to be .00 in both difficulty difference conditions by the authors. On the other hand, in the study that was carried out by Kim and Moses (2014), the difficulty difference of two conditions was set to be 0.70 for both easy and difficult modules.

In this study, in which a three-stage test is constructed, considering the fact that the difficulty values between the modules can increase in line with the number of stages (Schnipke & Reise, 1997), the difference between difficulty was set to be 0.5 in both 2. and 3. stages for small and large difficulty difference conditions. Moreover, the difficulty difference between easy modules (or difficult modules) was set to be 0.5 in the 2. and 3. Stages for two conditions. For 1-3-3 design; under the small-difference difficulty condition, the average of item difficulty parameters was set to be .00 for routing; for the second stage, -0.5 for easy, .00 for medium and +0.5 for difficult; for the third stage, -1.00 for easy, .00 for medium and +1.00 for difficult. Under large b parameter difference, the average of item difficulty parameters was set to be .00 for routing; for the second stage, -1.00 for easy, .00 for medium and +1.00 for difficult; for the third stage, -1.5 for easy, .00 for medium and +1.5 for difficult. As a consequence, when small b parameter difference condition is in question, the difference of range of parameters in the second stage is set to be 1.00, while it is set to be 2.00 in the third stage. In the case of large b parameter difference, on the other hand, the difference of range of b parameters is set to be 2.00 in the second stage, whereas it is set to be 3.00 in the third stage. For 1-2-2 panel design, the same item pool has been used while only the module with an average difficulty level at the second and third stage has been removed. For example, routing module with a module length of 10 items has been used for both small and large difference in 1-3-3 and 1-2-2 panel design. The easy and hard modules included in second stage of 1-3-3 panel design are common with the easy and hard modules included in 1-2-2 panel design under all conditions. The variables that are included in the study are summarized in [Table 1](#).

Table 1. *The variables that are included in the study*

Variable	Levels
Panel Design	“1-3-3”; “1-2-2”
Module Length	10-15-20
a parameter (item discrimination) sequence in stages	C1(“0.80”-“0.80”-“0.80”) C2(“1.40”-“0.80”-“0.80”) C3 (“0.80”-“1.40”-“0.80”) C4(“0.80”-“0.80”-“1.40”) C5 (“1.40”-“1.40”-“1.40”)
b parameter (difficulty) difference condition in stages	Small differences (1.00 theta differences in stage two; 2.00 theta differences in stage three) Large differences (2.00 theta differences in stage two; 3.00 theta differences in stage three)

2.2. Data Simulation

MSTGen (Han, 2013) was used for ca-MST application within the context of variables that are given in Table 1. More than one simulation were realized within the scope of the conditions specified in the program while constructing each module, and then, test information function (TIF) graphics were examined for that module before including the most suitable module according to the specified values in the scope of the study. For example, for 1-3-3- panel design small b difference condition; routing module is constructed in a way to reflect one TIF center (theta point of 0.00), the second stage is constructed in a way to reflect three TIF centers (theta points of -.05, .00, +0.5) and the third stage is constructed in a way to reflect three TIF centers (theta points of -1.00, .00, +1.00). 5000 simulees derived from normal distribution (N (0,1)) are simulated in this study. Maximum Fisher Information (MFI) module selection method was used to choose the modules. The method of Expected a Posteriori (EAP) was preferred for ability estimation of examinees. Moreover, 100 replications were carried out.

2.3. Data Analysis

In the study, only one panel implementation was realized while 60 conditions (2 panel designs \times 3 module lengths \times 5 item discrimination sequences \times 2 b parameter differences) were examined. For all conditions, Pearson product-moment correlation coefficient, RMSE (Root Mean Square Error) and AAD (Average Absolute Difference) were calculated. Also, the equations of Pearson product-moment correlation coefficient, RMSE and AAD are presented below.

$$r_{\hat{\theta}_i \theta_i} = \frac{n \sum_{i=1}^n \hat{\theta}_i \theta_i - \sum_{i=1}^n \hat{\theta}_i \sum_{i=1}^n \theta_i}{\sqrt{\left[n \sum_{i=1}^n \hat{\theta}_i^2 - \left(\sum_{i=1}^n \hat{\theta}_i \right)^2 \right] \left[n \sum_{i=1}^n \theta_i^2 - \left(\sum_{i=1}^n \theta_i \right)^2 \right]}}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (\hat{\theta}_i - \theta_i)^2}{n}}, \quad AAD = \frac{\sum_{i=1}^n |\hat{\theta}_i - \theta_i|}{n}$$

What each symbol represents in the formulas is given below.

n = the size of the sample

$\hat{\theta}_i$ = estimated level of ability for person i

θ_i = the known level of ability person i

When calculating Conditional RMSE, the groups were initially formed in each theta range of theta ability level and six groups were obtained from -3 theta to 3 theta values. Then, RMSE values of six theta θ change points were calculated.

3. RESULT / FINDINGS

The findings are given under two headings. Under the heading of overall outcomes, there are some explanations regarding goodness of fit values given in Table 2. Under the heading of conditional outcomes, graphs and explanations regarding the change of RMSE according to theta change points are presented.

3.1. Overall Outcomes

Correlation, RMSE and AAD values that have been obtained in relation to the 60 conditions that have been addressed within the framework of this study are given in Table 2.

Table 2. Corralation, RMSE and AAD results of ability estimation

Panel Design	b-parameter (difficulty) difference	a parameter(item discrimination) sequence*	Correlation			RMSE			AAD		
			Module Length			Module Length			Module Length		
			10	15	20	10	15	20	10	15	20
1-3-3	Small	C1	0.95	0.97	0.97	0.32	0.26	0.24	0.25	0.21	0.19
		C2	0.95	0.97	0.98	0.30	0.24	0.21	0.24	0.19	0.16
		C3	0.96	0.97	0.98	0.28	0.23	0.20	0.22	0.18	0.16
		C4	0.97	0.98	0.98	0.26	0.23	0.19	0.20	0.18	0.15
		C5	0.98	0.98	0.99	0.22	0.19	0.16	0.17	0.15	0.13
	Large	C1	0.95	0.97	0.97	0.32	0.26	0.24	0.25	0.21	0.19
		C2	0.96	0.97	0.98	0.30	0.23	0.21	0.23	0.19	0.16
		C3	0.97	0.98	0.98	0.26	0.21	0.19	0.20	0.17	0.15
		C4	0.97	0.98	0.98	0.27	0.22	0.21	0.21	0.18	0.17
		C5	0.98	0.98	0.99	0.22	0.18	0.16	0.17	0.14	0.13
1-2-2	Small	C1	0.95	0.96	0.97	0.32	0.27	0.24	0.25	0.21	0.19
		C2	0.96	0.97	0.98	0.30	0.24	0.21	0.24	0.19	0.17
		C3	0.96	0.97	0.98	0.29	0.23	0.21	0.23	0.18	0.16
		C4	0.97	0.97	0.98	0.27	0.23	0.20	0.21	0.18	0.16
		C5	0.97	0.98	0.99	0.23	0.19	0.16	0.18	0.15	0.13
	Large	C1	0.95	0.96	0.97	0.32	0.27	0.25	0.25	0.22	0.20
		C2	0.96	0.97	0.98	0.30	0.24	0.21	0.24	0.19	0.17
		C3	0.96	0.97	0.98	0.27	0.23	0.20	0.21	0.18	0.16
		C4	0.96	0.97	0.98	0.29	0.24	0.21	0.23	0.19	0.17
		C5	0.97	0.98	0.99	0.23	0.19	0.17	0.19	0.15	0.13

*C1(“0.80”-“0.80”-“0.80”) C2(“1.40”-“0.80”-“0.80”); C3(“0.80”-“1.40”-“0.80”); C4(“0.80”-“0.80”-“1.40”); C5(“1.40”-“1.40”-“1.40”)

The findings related to the values of goodness of fit that are given in [Table 2](#) are presented below.

3.1.1. When the design was 1-3-3 and there was a small difference in b parameters;

i) In ca-MST test structure whose module length was composed of 10 items, the lowest correlation value was found to be 0.95 in the case of C1 where item discrimination was selected to be the lowest across all stages. Similar to this result, the highest RMSE value (0.32) as well as the highest AAD value (0.25) were obtained. On the other hand, the highest correlation value of 0.97 was observed in the case of C5 condition where the items in all modules had the highest mean of item discrimination. The lowest RMSE value (0.22) besides the lowest AAD value (0.17) was obtained in this condition. On the other hand, among the conditions where the values of item discrimination were altered in stages, the highest correlation value (0.97), the lowest RMSE value (0.26) and the lowest AAD value (0.20) were obtained in the case of C4 condition where the items included in the last stage had the highest level of mean discrimination value.

ii) The lowest correlation value (0.97) as well as the highest RMSE value (0.26) and AAD value (0.21) in C1 condition were obtained under the condition where the module length covered 15 items. The highest correlation value (0.98) as well as the lowest RMSE value (0.19) and AAD value (0.15) were obtained under the condition of C5, where the items in all stages had the highest average value of item discrimination. On the other hand, when the change in item discrimination was considered, the highest correlation value (0.98) as well as the lowest RMSE value (0.23) and AAD value (0.18) were obtained under the condition of C4 where the items in the last stage had the highest value of item discrimination.

iii) In the test structure where the module length consisted of 20 items and when it was decided to have a small difference between b parameters as was the case when the module length was composed of 10 and 15 items, the lowest correlation value (0.97) as well as the highest RMSE (0.24) and AAD value (0.19) were obtained under C1 condition. The highest value of correlation (0.99), the lowest value of RMSE (0.16) and AAD (0.13) were obtained under C5 condition. When the change in item discrimination was considered, there was the highest value of correlation (0.98) as well as the lowest value of RMSE (0.19) and AAD (0.15) under the condition of C4 where the items included in the last stage had the highest value of item discrimination.

3.1.2. When the design was 1-3-3 and there was a large difference in b parameters

i) When the module length was decided to cover 10 items, C1 condition gave the lowest correlation value (0.95) while C5 condition gave the highest correlation value (0.98). In relation to the obtained correlation values, the highest RMSE value (0.32) and the lowest AAD value (0.25) were observed under C1 condition. On the other hand, the lowest RMSE value (0.22) and the lowest AAD value (0.17) were obtained in the case of C5 condition where item discrimination values were chosen to be equal and the highest in all stages. When the change in item discrimination values across stages was observed, the highest level of measurement precision was obtained under the condition of C3. The highest correlation value (0.97) as well as the lowest RMSE (0.26) and the lowest AAD value (0.20) were obtained under the condition of C3 where average value of item discrimination was chosen to be the highest in the second stage.

ii) Under the condition where the module length was composed of 15 items the lowest correlation value (0.97), the highest RMSE value (0.26) and AAD value (0.21) were obtained when a parameter sequence was C1. On the other hand, the highest correlation value (0.98) as well as the lowest RMSE value (0.18) and AAD value (0.14) were obtained in C5 sequence. Under the conditions related to the change in the values of item discrimination across stages,

measurement precision was found to be the highest under C3 condition. C3 condition produced a correlation value of 0.98, RMSE value of 0.21 and AAD value of 0.17.

iii) On the other hand, under the condition where the module length was composed of 20 items, the lowest value of correlation (0.97), the highest value of RMSE (0.24) and the highest value of AAD (0.19) were obtained in C1 sequence. Contrary to this condition, the highest value of correlation (0.99), the lowest value of RMSE (0.16) and the lowest value of AAD (0.13) were obtained under the condition of C5. When the change in the values of item discrimination across stages was considered, the highest value of measurement precision was obtained under C3 condition with the highest value of correlation (0.98), the lowest value of RMSE (0.19) and the lowest value of AAD (0.15).

Furthermore, when it comes to 1-3-3 panel design, it is observed that the more the module length increases, the more the correlation values increase in all different sequences of a parameters for both small difference and large difference conditions. When item discrimination sequence conditions are examined, the lowest goodness of fit values was obtained under C1 conditions whereas the highest values were obtained under C5 condition regardless of module length and item difficulty difference. When the sequences in which the average a parameter distribution showed variation across stages were examined, the highest level of measurement precision and a small difficulty difference were obtained under the condition of C4, whereas large difficulty difference was obtained under the condition of C3. In the tests with the same module length and with the conditions of both small difficulty difference and large difficulty difference, under the condition of C2, where the value of a parameter was chosen to be the highest in routing module, the measurement precision was found to be the lowest.

3.1.3. When the design was 1-2-2 and there was a small difference in b parameters

i) When the module length was composed of 10 items, the lowest value of correlation (0.95), the highest value of RMSE (0.32) and the highest value of AAD (0.25) were obtained under the condition of C1. On the other hand, the lowest value of RMSE (0.23) and the lowest value of AAD (0.18) were obtained when it comes to C5 condition, where the average a parameter values were equal and the highest, the highest value of correlation (0.97). On the other hand, when it comes to the conditions where the values of item discrimination were altered across stages, the condition of C4 had the highest measurement precision with the highest value of correlation (0.97), the lowest value of RMSE (0.27) and the lowest value of AAD (0.21).

ii) When the module length was composed of 15 items and there was a small difficulty difference, the lowest level of correlation (0.96) as well as the highest value of RMSE (0.27) and AAD (0.21) were obtained under the condition of C1. On the other hand, the highest value of correlation (0.98), the lowest value of RMSE (0.19) and the lowest value of AAD (0.15) were obtained under the condition of C5. When the change in item discrimination is considered, the condition of C4 produced the highest value of correlation (0.97) as well as the lowest value of RMSE (0.23) and the lowest level of AAD (0.18).

iii) When the module length was selected to be composed of 20 items, the lowest value of correlation (0.97) besides the highest value of RMSE (0.24) and the highest value of AAD (0.19) were obtained under C1 condition. In the condition of C5, where average a parameter value was chosen to be the highest in all modules, there came out the highest value of correlation (0.99), the lowest value of RMSE (0.16) and the lowest value of AAD (0.13). These results were similar to the results of those conditions where module lengths were chosen to be 10 and 15 items respectively. Furthermore, similar to the other module lengths, C4 condition, where the last stage included items with high values of item discrimination, produced the highest value of correlation (0.98), the lowest value of RMSE (0.20) and the lowest value of AAD (0.16).

3.1.4. When the design was 1-2-2 and there was a large difference in b parameters

i) When the module length was composed of 10 items and there were large difficulty difference conditions, the results were found to be similar to those that were obtained in the case of small difficulty difference conditions. The lowest value of correlation (0.95) and the highest value of RMSE (0.32) as well as the highest value of AAD (0.25) were obtained under the condition of C1. C5 condition, on the other hand, produced the highest value of correlation (0.97), the lowest value of RMSE (0.23) and the lowest value of AAD (0.19). The highest measurement precision in the change of values related to item discrimination across stages was observed in the condition of C3. C3 condition produced a correlation value of 0.96, RMSE value of 0.27 and AAD value of 0.21.

When the module length was composed of 15 items, the lowest value of correlation (0.96) as well as the highest RMSE value (0.27) and AAD value (0.22) were obtained under C1 condition. On the other hand, C5 condition resulted in the highest correlation value (0.98), the lowest RMSE value (0.19) and the lowest AAD value (0.15). The highest measurement precision was obtained under C3 condition in the change of item discrimination values across the stages. C3 condition produced a correlation value of 0.97, RMSE value of 0.23 and AAD value of 0.18.

When the module length was selected to be composed of 20 items and there was a large difficulty difference, the lowest value of correlation (0.97) besides the highest value of RMSE (0.25) and the highest value of AAD (0.20) were obtained under C1 condition. Contrary to this condition, the highest value of correlation (0.99) as well as the lowest value of RMSE (0.17) and the lowest value of AAD (0.13) were obtained under C5 condition. In C3 condition, where the value of average item discrimination was chosen to be highest in the second stage, there came out the highest value of correlation (0.98), the lowest value of RMSE (0.20) and the lowest value of AAD (0.16).

When 1-2-2 panel design was in question, it was observed that measurement precision increased with the increase in the module length for both small difficulty difference and large difficulty difference cases, which meant that the obtained goodness of fit values got better. When item discrimination sequence conditions are examined in general, the lowest goodness of fit value in C1 condition and the highest goodness of fit value in C5 condition were obtained regardless of module length and item difficulty difference.

In 1-2-2 panel design, when small difficulty difference condition was in hand, the highest value of measurement precision was obtained under C4 condition where the items having higher values of item discrimination were included in the last stage. However, when there was a large difficulty difference, higher values of goodness of fit were obtained under C3 condition where items having high values of item discrimination in all module lengths were included in the medium stage, when compared to the conditions where items with high values of item discrimination were included in the first and last stages. This was similar to the results obtained with 1-3-3 panel design.

Considering the effect of panel design, when all the other conditions are compared to each other, it is clear in some conditions that the correlation values are higher whereas error values are lower in 1-3-3 panel design in each of the conditions included in this study.

3.2. Conditional Outcomes

Conditional RMSE values are examined at six theta change points within the framework of this study. Figure 2 below shows the conditional RMSE values.

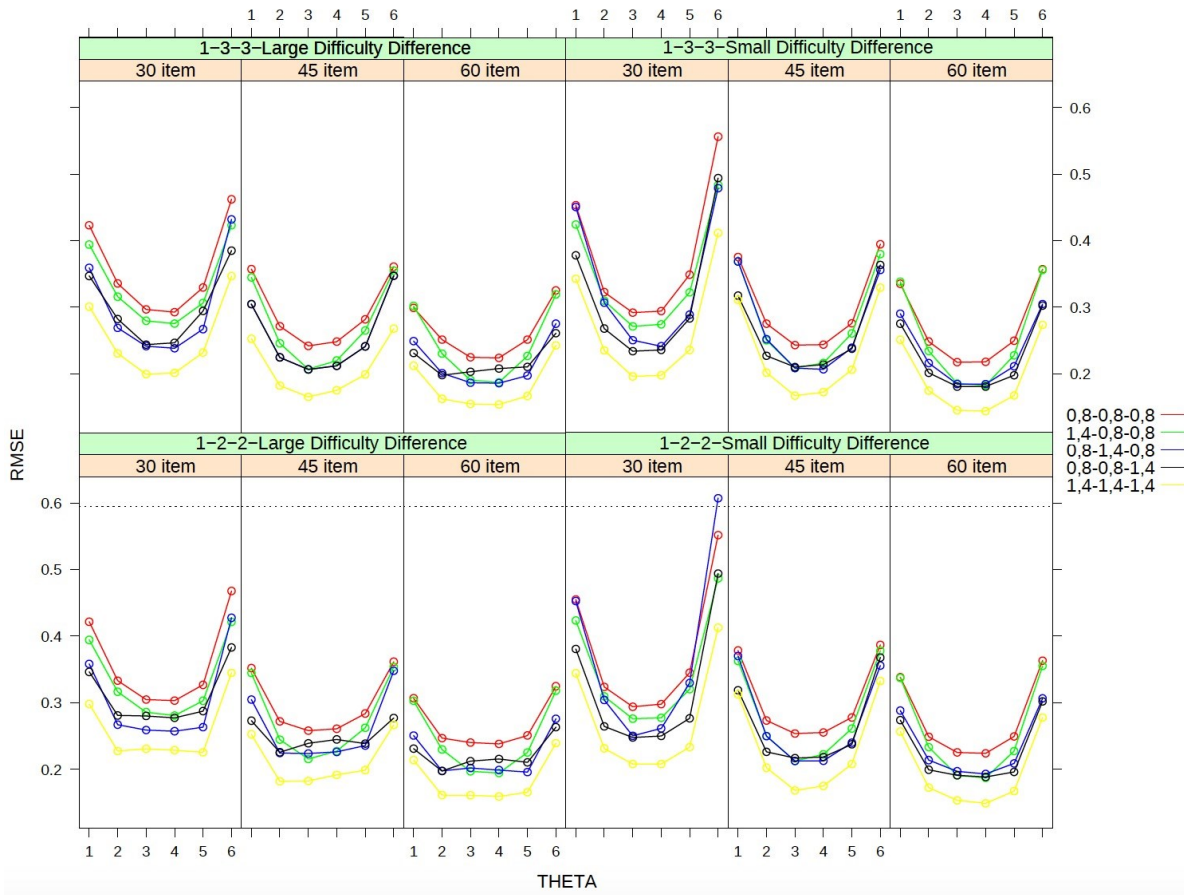


Figure 2. Conditional RMSE's of All Test Designs

When Figure 2 is examined, it is obvious that distribution of errors is lower at the level of extreme ability in all cases when compared to the ones at the level of medium ability. It is also observed that errors at all ability levels decrease with the increase in the module length. Especially in the module lengths with 20 items, it is seen that the difference between the errors at extreme and medium level abilities decrease. When the change of RMSE values in small and large b difference conditions are observed, it can be stated that errors tend to be higher at extremely high or low ability levels in small difficulty difference conditions with the same module length. These results are also valid for both 1-3-3 and 1-2-2 panel designs.

Moreover, at all ability levels, the distribution of errors in general are lower under C5 condition, where a parameter values are chosen to be higher in all stages. However, as it is seen in Figure 2, RMSE values at extreme ability levels are generally observed to be closer to each other under C1 and C2 conditions. Also, when the results obtained from C3 and C4 conditions are examined, close RMSE values are obtained at extreme ability levels.

4. DISCUSSION

The effect of different panel designs, module lengths, different sequence of a parameter value across stages and change in b parameter range on measurement precision in ca-MST implementations have been investigated within the scope of this study. The values of correlation, RMSE and AAD for 60 conditions addressed for that purpose have been calculated.

When the effect of test length is examined, the research result showed that there occurs a decrease in RMSE values at all ability levels as the test length increases. There are studies in the literature that have obtained similar results. Kim and Plake (1993) carried out a study in which they found out that RMSE values decrease as the test length increases in two-stage tests that are composed of 40-45-60 items, respectively. Within the framework of this study, all the modules have been designated to include equal numbers of items. This means that there is an increase in the number of items included in routing module as well as the following modules as the test length increases. The studies that focus on the length of routing module in the literature have given similar results to this study, which puts forth that errors decrease as the test length increases (Kim & Plake, 1993; Kim et al., 2015; Loyd, 1984). Moreover, when conditional RMSE values that are obtained for each ability level are examined, it becomes clear that there are more errors in tests with lower values of test length at extreme ability levels, whereas measurement precision increases as the module length increases.

The study has also focused on investigating the difference between b parameters of modules. When the overall outcomes are examined, differences b parameter did not make any impact on the outcomes. In extremely high or low ability levels, the condition of small difficulty difference has produced higher levels of error irrespective of other conditions. Especially when the difficulty of modules in the second stage is closer to the difficulty level of routing module, poor measurement can be obtained for the individuals with extreme ability levels (Lord, 1971; Patsula, 1999). As a consequence, the condition of large difficulty difference can ensure a higher level of measurement precision when estimating the abilities of individuals with extreme levels of ability. Kim et al. (2015) carried out a study in which they investigated small and large difficulty difference conditions when various ability estimations are in question in two-stages tests. Similar to the results of this study, they have concluded that lower levels of error are obtained (in some ability estimations) under the condition of large difficulty difference in extreme ability level. Test developers can be recommended in the light of the study results to include very easy and/or very difficult items in the ca-MST item pool for the purpose of measurement precision. However, it can be difficult to develop very difficult test items when compared to easy items or the ones with a medium level of difficulty (Kim & Moses, 2014).

When the average values of item discrimination belonging to the items included in the modules are considered, it is obvious that the lowest error is gained under the condition of C5, where a parameter values at all stages is equal and the highest. It is an expected result to have more reliable measurement as item discrimination values increase. Another question which is discussed within the scope of this study is at which stage the items with higher values of item discrimination should be included in order to reduce the errors. In line with this question, an important result of the study is that a high degree of measurement precision is obtained when small difficulty difference condition is in hand under the condition of C4, where the items at the last stage have high values of item discrimination. Under the condition of large difficulty difference, on the other hand, a high degree of measurement precision is obtained under the condition of C3, where the medium stage consists of items with high values of item discrimination. At the same time, the difference the b parameters were chosen as 2.00 theta for both the last stage of the small difficulty difference condition and the second stage of the large difficulty difference condition. At the end of the study, it was also discovered that measurement precision does not increase even if items with high values of item discrimination are used when the difference between b parameters becomes larger.

When using items with high values of item discrimination in the routing module, second stage and last stage in terms of ability levels are considered and, it is observed that the individuals with medium levels of abilities get errors closer to each other under the three conditions (C2, C3 and C4). However, when the test is short, the condition of C2 gives high values of errors at

the medium ability level. When extremely high or low ability levels are in question, including items with high values of item discrimination in the medium and last stages gives similar results, whereas including these items in the last stage produces lower levels of errors. The results of this study are parallel with the results of the study carried out by Chang and Ying (1999) as well as Zheng et al. (2012). It can be recommended to test operators to make use of item pools with high values of item discrimination as it will increase measurement precision (Xing & Hambleton, 2004). However, it can be stated that when this condition cannot be ensured, including items with high values of item discrimination in the last stage can contribute to measurement precision. It can be inferred from the results of this study that including items with high values of item discrimination in the routing module does not have any impact on measurement precision.

When the impact of panel design is considered, measurement precision of 1-3-3 panel design is higher than that of 1-2-2 panel design in some conditions. This can be explained via the fact that the items that are appropriate for medium level of ability are included in 1-3-3 panel design. When there is an increase in the number of modules, measurement precision also increases (Patsula, 1999). However, when RMSE values obtained from both 1-3-3 and 1-2-2 panel designs are examined in terms of ability levels, the errors obtained at the medium ability level are fewer than the errors at the extreme ability levels.

It can be recommended to the researchers in light of the results of this study to carry out similar studies with different panel designs (1-2-4; 1-2-3; 1-2-3-4) including different modules or stages. The effect of content distribution is not addressed in this study, so the effect of conditions with contents of different weights can be investigated. The number of items included in the modules are fixed in this study. The future studies should have a new condition to include different number of items in the modules. Moreover, it can be suggested to analyze the effect of module selection methods.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author(s).

ORCID

Melek Gülşah Şahin  <https://orcid.org/0000-0001-5139-9777>

5. REFERENCES

- Boztunç Öztürk, N. (2019). How the length and characteristics of routing module affect ability estimation in ca-MST?. *Universal Journal of Educational Research*, 7(1), 164-170. <https://doi.org/10.13189/ujer.2019.070121>
- Chang, H., & Ying, Z. (1999). a-Stratified multistage computerized adaptive testing. *Applied Psychological Measurement*, 23(3), 211-222. <https://doi.org/10.1177/01466219922031338>
- Chen, L. Y. (2010). An investigation of the optimal test design for multi-stage test using the generalized partial credit model (unpublished doctoral dissertation). The University of Texas at Austin. Retrieved from <https://repositories.lib.utexas.edu/handle/2152/ETD-UT-2010-12-344>
- Hadadi, A., & Leucht, R. M. (1998). Some methods for detecting and understanding test speededness on timed multiple-choice tests. *Academic Medicine*, 73, 47-50. https://journals.lww.com/academicmedicine/Citation/1998/10000/TESTING_THE_TES_T_Some_Methods_for_Detecting_and.42.aspx

- Hambleton, R.K., & Xing, D. (2006). Optimal and nonoptimal computer-based test designs for making pass-fail decisions. *Applied Measurement in Education*, 19(3), 221-229. https://doi.org/10.1207/s15324818ame1903_4
- Han, K. T. (2013). MSTGen: simulated data generator for multistage testing. *Applied Psychological Measurement*, 37(8), 666-668. <https://doi.org/10.1177/0146621613499639>
- Han, K. T., & Guo, F. (2013). *An approach to assembling optimal multistage testing modules on the fly* (Report No. RR-13-01). Virginia: GMAC.
- Hendrickson, A. (2007). An NCME instructional module on multistage testing. *Educational Measurement: Issues and Practice*, 26, 44-52. <https://doi.org/10.1111/j.1745-3992.2007.00093.x>
- Jodoin, M. G., Zenisky, A., & Hambleton, R. K. (2006). Comparison of the psychometric properties of several computer-based test design for credentialing exams with multiple purposes. *Applied Measurement in Education*, 19(3), 203-220. http://doi.org/10.1207/s15324818ame1903_3
- Kim, H., & Plake, B.S. (1993, April). *Monte carlo simulation of two-stage testing and computerized adaptive testing*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Atlanta.
- Kim, J., Chung, H., Dodd, B.G., & Park, R. (2012). Panel design variations in the multistage test using the mixed-format tests. *Educational and Psychological Measurement*, 72(4), 574-588. <https://doi.org/10.1177/0013164411428977>
- Kim, S., Moses, T., & Yoo, H. (2015). A comparison of IRT proficiency estimation methods under adaptive multistage testing. *Journal of Educational Measurement*, 52(1), 70-79. <https://onlinelibrary.wiley.com/doi/abs/10.1111/jedm.12063>
- Kim, S., & Moses, T. (2014). An investigation of the impact of misrouting under two-stage multistage testing: A simulation study (Report No. RR-14-01). Princeton, NJ: English Testing Service.
- Leucht, 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.
- Leucht, R., Brumfield, T., & Brithaupt K. (2006). A testlet assembly design for adaptive multistage tests. *Applied Measurement in Education*, 19(3), 189-202. https://doi.org/10.1207/s15324818ame1903_2
- Luecht, R. M., & Nungester, R. J. (1998). Some practical examples of computer-adaptive sequential testing. *Journal of Educational Measurement*, 35(3), 229-249.
- Leucht, R. M., & Sireci, S. G. (2011). *A review of models for computer-based testing*. (Report No. RR-2011-12). New York: CollegeBoard. <https://doi.org/10.1111/j.1745-3984.1998.tb00537.x>
- Lord, F. M. (1971). A theoretical study of two-stage testing. *Psychometrika*, 36(3), 227-242. <https://doi.org/10.1007/BF02297844>
- Loyd, B. (1984, February). Efficiency and Precision in two-stage adaptive testing. Paper presented at the Annual Meeting of Eastern Educational Research Association, West Palm Beach, FL.
- Park, R. (2015). Investigating the impact of a mixed-format item pool on optimal test designs for multistage testing (Doctoral dissertation). The University of Texas at Austin. Retrieved from <https://repositories.lib.utexas.edu/handle/2152/31011>
- Patsula, L.N. (1999). A comparison of computerized adaptive testing and multistage testing. (Doctoral dissertation). The University of Massachusetts Amherst. Retrieved from

- https://scholarworks.umass.edu/cgi/viewcontent.cgi?article=4283&context=dissertations_s_1
- Rotou, O., Patsula, L., Steffen, M., & Rizavi, S. (2007, March). *Comparison of multistage tests with computerized adaptive and paper and pencil tests*. (Report No: RR-07-04). Princeton, NJ: English Testing Service.
- Sarı, H.İ., Yahşi Sarı, H., & Huggins Manley, A.C. (2016). Computer adaptive multistage testing: Practical issues, challenges and principles. *Journal of Measurement and Evaluation in Education and Psychology*, 7(2), 388-406. <https://doi.org/10.21031/epod.280183>
- Schnipke, D.L., & Reese, L.M. (1999). *A comparison of testlet-based test designs for computerized adaptive testing*. (Report No: 97-01). Princeton, NJ: Law School Admission Council.
- Sadeghi, K., & Khonbi, Z.A. (2017). An overview of differential item functioning in multistage computer adaptive testing using three-parameter logistic item response theory. *Language testing in Asia*, 7(1), 1-16. <https://doi.org/10.1186/s40468-017-0038-z>
- van der Linden, W.J. (2005). *Linear models for optimal test design*. New York: Springer.
- Verschoor, A., & Eggen, T. (2014) Optimizing the test assembly and routing for multistage testing. In D. Yan., A. A. von Davier, & C., Lewis, (Ed.), *Computerized Multistage Testing Theory and Applications* (pp:135-150). Taylor & Francis Group.
- Xing, D., & Hambleton, R., K. (2004). Impact of test design, item quality, and item bank size on the psychometric properties of computer-based credentialing examinations. *Educational and Psychological Measurement*, 64(1), 5-21. <https://doi.org/10.1177/0013164403258393>
- Zeng, W. (2016). Making test batteries adaptive by using multistage testing techniques (Doctoral dissertation). The University of Wisconsin-Milwaukee. Retrieved from <https://dc.uwm.edu/cgi/viewcontent.cgi?article=2241&context=etd>
- Zenisky, A.L., & Hambleton, R., K. (2014). Multistage test desing: Moving research results into practice. In D. Yan., A. A. von Davier, & C., Lewis, (Ed.), *Computerized Multistage Testing Theory and Applications* (pp. 21-37). Taylor and Francis Group.
- Zenisky, A., L., & Jodoin, M., G. (1999). Current and future research in multistage testing. (Report No:370). Amherst, MA: University of Massachusetts School of Education.
- Zenisky, A. L. (2004). Evaluating the effects of several multi-stage testing design variables on selected psychometric outcomes for certification and licensure assessment. (Doctoral dissertation). University of Massachusetts Amherst. Retrieved from <https://scholarworks.umass.edu/dissertations/AA13136800>
- Zheng, Y., Nozawa, Y., Gao, X., & Chang, H. (2012). *Multistage adaptive testing for a large-scale classification test: Design, heuristic assembly, and comparison with other testing modes*. (Report No:2012-6). Iowa City, IA.: ACT.
- Zheng, Y., & Chang, H. (2015). On-the-fly assembled multistage adaptive testing. *Applied Psychological Measurement*, 39(2), 104-118. <https://doi.org/10.1177/0146621614544519>