International Journal of Assessment Tools in Education
2024, Vol. 11, No. 2, 270–287

https://doi.org/10.21449/ijate.1268614

journal homepage: https://dergipark.org.tr/en/pub/ijate *Research Article*

Investigation of a multistage adaptive test based on test assembly methods

Ebru Doğruöz [1,*], Hülya Kelecioğlu [2]

[1]Çankırı Karatekin University, Faculty of Humanities and Social Sciences, Department of Educational Measurement and Evaluation, Çankırı, Türkiye
[2]Hacettepe University, Faculty of Education, Department of Educational Measurement and Evaluation, Ankara, Türkiye

ARTICLE HISTORY

Received: Mar. 21, 2023

Accepted: Feb. 13, 2024

Keywords:

Adaptive multistage test,
Test assembly,
PISA.

Abstract: In this research, multistage adaptive tests (MST) were compared according to sample size, panel pattern and module length for top-down and bottom-up test assembly methods. Within the scope of the research, data from PISA 2015 were used and simulation studies were conducted according to the parameters estimated from these data. Analysis results for each condition were compared in terms of mean RMSE and bias. According to the results obtained from the MST simulation based on the top-down test assembly method, mean RMSE values reduced when the module length increased and when the panel pattern changed from 1-2 to 1-2-2 and 1-2-3 for MST applied to small and large samples. Within the scope of the research, data from PISA 2015 were used and simulation studies were conducted using the parameters estimated from these data. Analysis results for each condition were compared in terms of mean RMSE and bias.

1. INTRODUCTION

The combination of computer technology and test implementations with item response theory (IRT) led to the emergence of computer adaptive tests (CAT). While these tests involve the use of a computer and are tailored to the examinee, IRT allows the opportunity to develop, apply and evaluate a test by considering the abilities of the examinee. Due to these advantages, CAT was used instead of paper and pencil tests. The first application of an adaptive test in the computer environment was completed by Reckase in 1974 (Wise & Kingsbury, 2000). In addition, the emergence and development of item response theory has enabled the realization of adaptive tests through the parameterization of examinee's abilities and item characteristics (Linden & Glas, 2000). Through computers, the examinees' ability can be estimated instantly after each response to an item. Thus, the next item is selected according to the examinee's ability. Accordingly, CAT has been adopted and used in many national and international exams around the world (Khorramdel et al., 2020; Kirsch & Lennon, 2017). Today, some of these exams prefer MST instead of CAT. For example, GRE (Graduate Record Examinations), PIAAC (Program for International Assessment of Adult Competencies), AICPA (American Institute of Certified Public Accountants') and MAPT (Massachusetts Adult Proficiency Test) use MST instead of CAT because of its advantages (American Institute of Certified Public Accountants, 2019; Educational Testing Service, 2018; Hogan et al., 2016; Zenisky et al.,

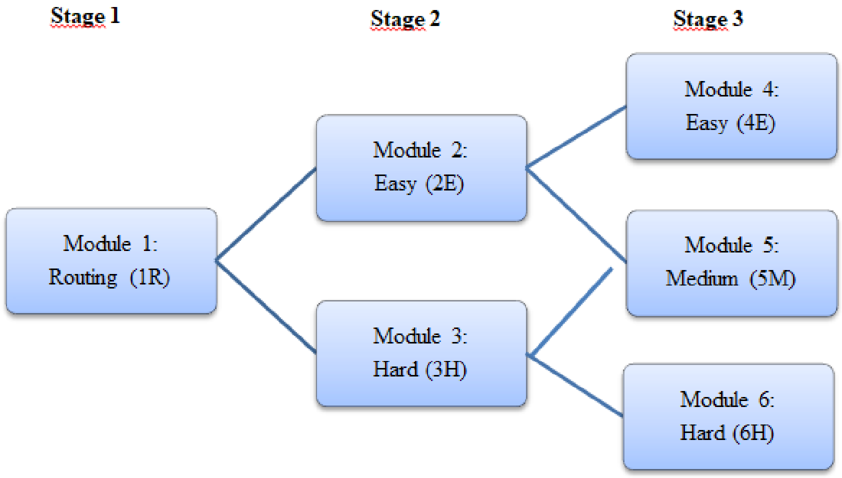
*CONTACT: Ebru Doğruöz ✉ ebrudemircioglu@karatekin.edu.tr Çankırı Karatekin University, Faculty of Humanities and Social Sciences, Department of Educational Measurement and Evaluation, Çankırı, Türkiye

e-ISSN: 2148-7456 /© IJATE 2024

2009). One of the reasons behind this trend is that MST acts as a bridge between linear test forms of paper and pencil testing and computer-based tests and computer-based test forms that are adaptable at item level. MST is both an adaptive test and also allows the opportunity for the test developer to investigate the test form ahead of time and check examinee's responses (Yan et al., 2014).

MST is defined as a a type of computerized adaptive testing allowing adaptation of the difficulty of the test according to the ability level of the examinee being tested. This assessment type comprises clustered components called *modules*, *stages*, *panels* and *pathways*. The smallest element of this cluster is the *module*. A module is a group of items formed by bringing items together. The level of module or modules is called the *stage*. A *panel* is a pattern formed by combining stages. The panel is the largest component of MST. For example, a panel formed with 1 module in the first stage, 2 modules in the second stage and 3 modules in the third stage is called the '1-2-3' MST panel pattern. The route taken by an examinee between stages and modules in the panel is called the *pathway*. Each examinee only follows one pathway during the test (Zenisky & Hambleton, 2014). The schematic appearance of the MST components is presented in Figure 1.

Figure 1. *An example of 3-stage MST panel.*



1.1. Test Assembly

Based on a variety of statistical features, the combination of items chosen from the item pool on test form is called test assembly. The assembly of the forms is formulated as a combinatorial optimization (CO) problem, referred to as the test assembly problem (Papadimitriou & Steiglitz, 1982; Theunissen, 1985; van der Linden & Boekkooi-Timminga, 1989). CO is the research of an element in a finite cluster optimized to a certain function. The CO problem may be formulated as in Equation 1.1:

$$\text{To maximize } \mathbf{F}(\mathbf{x}) \qquad \text{Subject to } \mathbf{x} \in X \tag{1.1}$$

$\mathbf{x} = (x_1, x_2, \ldots\ldots, x_n)^T$ is a binary decision vector describing a test. When $x_i = 1$, the item i is included in the test; when $x_i = 0$ the item i is not included on the test.

n is the number of items in the item pool.

X includes all binary vectors each describing a feasible test. For this reason, this set is called the *feasible set*. In practice, the feasible set is not given explicitly; however, it is implicitly indicated by an equation constraining the decision vector and a list of inclusions. This list directly comprises the test properties. For example, the applicable set containing items from 5 to 10 is presented in Equation 1.2:

$$5 \leq \sum_{i=1}^{n} x_i \leq 10 \tag{1.2}$$

$$x_i \in \{0,1\}$$

For this feasible set, the second restriction does not involve any CO problem. For example, for each appropriate solution **x** = $(x_1, x_2, \ldots\ldots, x_n)^T$ there should be a binary vector.

F(x) is a vector function; in other words, the target function (Veldkamp, 1999). For example, the Maximum Fisher Information of an adaptive with θ' ability estimation is calculated with the function in Equation 1.3:

$$\text{To maximize } \sum_{i=1}^{n} \mathrm{I_i}(\theta'\,)x_i \tag{1.3}$$

$\mathrm{I_i}(\theta')$ is the Fisher information for item *I* at θ' ability level (Lord, 1980).

Accordingly, estimating the maximum number of non-overlapping tests that can be obtained from an item pool given the test characteristics is very important in the construction of the item pool. It should be noted that test pooling for MST is a very complex process. This is because test combination in MST is realized by simultaneously creating many panels that are parallel in terms of both coverage and psychometric properties. This combination is performed in two steps: (1) assembling modules from the item pool and (2) assembling panels of modules. These panels should also consist of modules that fulfill certain statistical requirements, such as target test information functions (TIFs) (Luecht & Nungester, 1998). In this context, limitations related to content balancing, exposure control, coverage effects, cognitive knowledge levels of test takers, item and test item overlap, item format, and word count must also be met (Hendrickson, 2007). For this reason, test combining in MSTs is usually performed through automatic test assembly (ATA) algorithms and computer programs (Breithaupt & Hare, 2007; Breithaupt et al., 2005; Luecht, 2000; Luecht, 2006; Luecht et al., 2006; Luecht & Nungester, 1998; van der Linden, 2005).

1.1.1. *Automated test assembly method*

Automated test assembly (ATA) is a modern approach to test assembly that applies advanced optimization algorithms on computers to automatically generate test forms. The most important feature of ATA is that it greatly improves the efficiency and accuracy of test assembly. This is because ATA enables computer-based selection of a suitable set of items from a large pool of pre-calibrated items (Theunissen, 1985; van der Linden, 2005; van der Linden & Boekkooi-Timminga, 1989; Veldkamp et al., 2013). The automated test assembly method may be applied with ATA computer software (e.g., CASTISEL, ConTEST) making calculation processes easier for test developers. The aim is to create test panels by choosing items from the item pool in modules taking into account the constraints such as content area, word count and item type. In this way, the process of choosing items from the item pool for modules is more convenient. This situation allowed the module development process to become more standardized.

1.1.2. *Test assembly methods: Top-down and bottom-up*

Luecht and Nungester (1998) recommended two strategies for the assembly of MST panels: top-down test assembly and bottom-up test assembly. Both strategies first require items to be assembled one by one into modules, then modules are assembled into panels. However, there are statistical differences in the stage of creating panels by combining modules between the strategies. The top-down test assembly strategy freely mixes and matches modules to create panels. The bottom-up test assembly strategy requires selective matching of modules to create panels. This is an indicator that the top-down test assembly method has a more complicated structure compared to the bottom-up test assembly method.

When combining modules to create panels in both test assembly strategies, the following steps are taken (Luecht & Nungester, 1998):

a) Production of statistical targets for test samples in different stages,

b) Determination of content features in the stages,
c) Creation of panels by combining modules abiding by the restrictions in the first and second steps.

Selection of statistical targets for modules is the most important decision in designing the MST pattern (Hendrickson, 2007). Zheng et al. (2012) created MST according to the top-down test assembly method based on the automatic approach. They compared this method with paper and pencil testing and CAT. According to the results of the study, MST utilized the item pool more effectively compared to pencil paper test and CAT, and the classification was performed more accurately. In a study discussing possible applications of adaptive or multistage tests for a Law Faculty Acceptance Test and considering the main approaches applied in the development of test assembly methods, a single-form test assembly approach was concluded to be an applicable method for testing in programs where the test is defined only by restrictions (Belov, 2016).

When research about test assembly methods is generally investigated, the common point appears to be that studies researched the top-down test assembly method proposed by Luecht and Nungester (1998) among test assembly methods and the test assembly method completed during exams. However, there are no experimental studies on how these test assembly methods give results under different conditions. Therefore, there is a question mark about whether the right decision is made in determining the test assembly method to be selected. The bottom-up test assembly method is chosen less often than the top-down test assembly method and is more advantageous for short test applications, which has made the top-down test assembly method a focal point for research. In the related literature, the bottom-up test assembly method was mostly used in the existing applications of MST (Hembry, 2014; Jodoin et al., 2006; Lu, 2010; Luecht et al., 2006; Wang, 2013; Wang, 2017; Yang, 2016; Zheng, 2014). There are a few studies in which the top-down test assembly method was used (Davis & Dodd, 2003; Lynn Chen, 2010; Zheng et al., 2016). For this reason, it is believed that this study can guide researchers on which of the 'top-down' or 'bottom-up' test assembly methods to prefer in the process of constructing the MST. Additionally, comparisons were made of the elements comprising MST like panel pattern, module length and stage number. In this framework, recommendations were developed regarding the module length, panel pattern and sample size required to make estimations with minimum error and bias in MST constructs. Because measurement precision in MSTs can be affected by module length and panel pattern (Zenisky & Hambleton, 2014). In addition, within the scope of the study, data from PISA 2015 were used and a simulation study was conducted using the parameters estimated from these data. PISA 2015 is an international, validated and reliable assessment, and this computer-based application is the basis for the MST to be used in the coming years, which is one of the reasons why PISA data were preferred in the research. Thus, a post-hoc simulation study was conducted based on real data. This is one of the important features that make the research strong. The results obtained in the study are expected to contribute to the applicability of MST. In line with this, within the scope of the present research, the aim was to compare test assembly methods and answer the following questions.

How do test assembly methods (top-down, bottom-up) impact the estimation of ability estimation conditional on module length, panel pattern, and sample size?

1.2. Subproblems

1. What changes occur in RMSE and bias values according to module length (6 and 12), panel pattern (‘1-2’, ‘1-2-2’ and ‘1-2-3’) and sample size (250 and 2000) for the top-down test assembly method in MST applications?
2. What changes occur in RMSE and bias values according to module length (6 and 12), panel pattern (‘1-2’, ‘1-2-2’ and ‘1-2-3’) and sample size (250 and 2000) for the bottom-up test assembly method in MST applications?

2. METHOD

2.1. Research Model

In this research, the aim was to compare the performance of test assembly methods for MST patterns with different features using IRT-based estimation and post-hoc simulation methods for the science literacy ability of examinees participating in the PISA implementation completed in 2015. For this purpose, real item data were used in the study. Therefore, this study is descriptive research based on post hoc simulation using real item parameters. Simulation studies consist of data generation and analysis processes appropriate to real-life situations (Burton et al., 2006; Ranganathan & Foster, 2003). Simulation data are often preferred because most MST applications have implementation problems, require a large sample size and a large item bank (Pihlainen et al., 2018; Xu et al., 2021; Zheng & Chang, 2015).

2.2. Participants

The participants in the research comprised examinees participating in PISA 2015 in the field of science literacy and answered booklet 91 because it is suitable for the structural features of the MST. The reason for choosing science literacy is that it constitutes the predominant area of the PISA 2015 application. Nearly 540,000 students representing 29 million students in nearly 72 countries, including 35 OECD countries, participated in the PISA 2015 implementation (OECD, 2015). Booklet number 91 was chosen as the data collection tool for the study, since the number of science literacy items and examinees who received the booklet were higher than for the other booklets, out of a total of 66 booklets (Forms 31-96) created according to the computer-based test. This booklet contained a total of 501 items in a variety of categories (two categories, multiple categories, open-ended) in the science literacy field. The item pool for the study comprised 159 items with two categories among the total of 501 items in booklet number 91. Analyses were completed on the dataset related to 15,059 students who answered these items.

2.3. Analysis

Analysis of data in the study was completed in two stages. In the first stage, data obtained from the study group comprising students participating in the PISA exam in 2015 were analyzed according to the 2 PL model based on IRT and an item pool was created for MST. In line with this, first, the data set obtained from the PISA implementation in 2015 was tested for a single dimension, local independence, model-data fit, item and ability parameter invariance assumptions. The suitability of the data set for factor analysis was tested with Bartlett's test and Kaiser-Meyer-Olkin (KMO) criteria (Bartlett's = 1584902.1, *sd* = 12561, *p* = 0.00; KMO = 0.98) and the data set was suitable. Item parameters and ability parameters related to examinees were estimated with the BILOG-MG (Zimowski et al., 1996) program. As the items had low correlation in the limited ability interval and the single dimension assumption was met, the local independence assumption was accepted. With the aim of investigating which logistic model was suitable for the data set, the data set was analyzed for suitability to 1 PL, 2 PL and 3 PL models. Accordingly, considering the difference between the –2 log (probability) values for 3 PL and 2 PL models was not much, the 2 PL model was chosen (–2 log (probability) $_{(1\ \mathrm{PL})}$ = 2125726.00 –2 log (probability) $_{(2\ \mathrm{PL})}$ = 2017798.00, (–2 log (probability) $_{(3\ \mathrm{PL})}$ = 1977773.91). These results are consistent with the technical report released by OECD (OECD, 2017). Thus, the 2 PL model was estimated to be suitable for calibration of the two-category data set identified to have a single dimension. According to descriptive statistics related to item and ability parameters, the data set had an item discrimination parameter value mean 1.16 and a standard deviation of 0.06 and a difficult parameter value mean 0.07 and a standard deviation of 0.30. The smallest ability parameter of individuals was -2.85, while the highest ability parameter was calculated as 2.97. In order to determine the invariance of item parameters, individuals were randomly divided into 11 groups. The item parameters were estimated according to the 2 PL model in different groups and the item parameters were compared between groups with the

Pearson moment multiplication correlation technique. Finally, significant and high levels of correlation (*p*<0.01) were identified between item parameters estimated in 11 groups comprising 1.369 individuals each and the invariance of item parameters assumption was met. The invariance of ability parameters was identified with significant positive, high-level correlations between ability parameters estimated in three randomly assigned subgroups comprising 53 items for all 15.059 individuals.

In the second stage of data analysis, an MST simulation was developed for each subproblem. Analyses were completed with item parameters chosen in accordance with 24 simulation conditions from the item pool and according to individual ability chosen in accordance with 24 simulation conditions. To create the MST, the 'xxIRT' (Luo, 2017) program using R (R Development Core Team, 2011) software was used. With the aim of increasing the generalizability of the results, 30 repeats were performed for each condition (Tian, 2018). The MST variables used in the MST simulations were test assembly (top-down and bottom-up), module length (6 and 12), panel pattern ('1-2', '1-2-2' and '1-2-3') and sample size (250 and 2000).

2.3.1. *Panel pattern*

In the research, MSTs with two ('1-2') and three ('1-2-2' and '1-2-3') stage panel patterns were created. These three-panel patterns were used in the research as they are included among the most researched MST panel patterns (Jodoin et al., 2006; Luecht et al., 2006; Wang, 2017; Zenisky, 2004). The '1-2' panel pattern comprises two stages and one panel. In the first stage, there is Module-1 (M) with a moderate difficulty level and in the second stage there is Module-2 (E) with an easy difficulty level and Module-2 (H) with a high difficulty level. The '1-2-2' panel pattern comprises three stages and two panels. The first stage includes Module-1 (M) with moderate difficulty, the second stage includes Module-2 (E) with easy difficulty and Module-2 (H) with high difficulty level and the third stage includes Module-3 (E) with easy difficulty and Module-3 (H) with high difficulty level. The '1-2-3' panel pattern comprises three stages and two panels. The first stage includes Module-1 (M) with a moderate difficulty level, the second stage comprises Module-2 (E) with easy difficulty and Module-2 (H) with high difficulty and the third stage includes Module-3 (E) with easy difficulty, Module-3 (M) with moderate difficulty and Module-3 (H) with high difficulty level.

2.3.1.1. Module length. The test length in MST studies was identified to vary between 33 and 60 items (Hambleton & Xing, 2006; Jodoin et al., 2006; Patsula, 1999; Zenisky, 2004). In this research, the number of modules representing short test length was chosen as 6, while the number of modules representing moderate test length was determined to be 12, twice that of the short test length. MSTs were designed so that when module length was 6, with the '1-2' panel pattern, individuals answered a total of 12 items, and with the '1-2-2' and '1-2-3' panel patterns they answered a total of 18 items. When the module length was 12, with the '1-2' panel pattern, individuals answered a total of 24 items, and with the '1-2-2' and '1-2-3' panel patterns they answered a total of 36 items.

2.3.1.2. Item Pool. There was a total of 159 items in the two-category data set calibrated according to the 2 PL model obtained from the PISA data administered in 2015.

2.3.1.3. Sample Size. The research sample comprised 250 and 2000 individuals chosen at random from among 15,059 individuals participating in the PISA test in 2015. When the literature is investigated, it appears sample sizes from 250 (Yan et al., 2014) to 5000 studies in MST research (Dallas, 2014; Sari, 2016; Wang, 2017; Xing & Hambleton, 2004; Yang, 2016). In this research, as the target was to assess the applicability to small samples in addition to large samples for the MST pattern, in the research 250 individuals represented the small sample and 2000 individuals represented the large sample.

2.3.1.4. Test Assembly. Most commonly used top-down and bottom-up automated test assembly methods used in MST studies were chosen. For both methods, the target test information function (TIF) value was determined with the mean maximum information (MMI) (Luecht, 2000; Luecht et al., 2006) strategy.

2.3.1.5 Referral Strategy and Scoring. In this study, the referral strategy was chosen as the commonly-used mean maximum information (MMI) strategy and scoring was done according to maximum likelihood estimation (MLE) (Luecht et al., 2006; Zenisky et al., 2010).

2.3.2. *Test administration*

The steps followed during test administration of the '1-2', '1-2-2' and '1-2-3' panel patterns investigated in the study were: (a) the individual was assigned one of two different panel patterns at random, (b) the individual responded the referral module (moderate difficulty) they were assigned, (c) after completing the referral module, the individual's ability was estimated using the maximum probability estimation (MPE), (d) after the first stage, the estimated ability of the individual (θ) and previously determined referral points were compared, and the individual was directed from the first stage to the second stage, (e) after the second stage, the individual's ability was again estimated with the MPE method, and (f) the test ended here for individuals tested with the '1-2' panel structure. For test administration of individuals tested with the '1-2-2' and '1-2-3' panel structures, their ability was predicted after the second stage (θ) and compared with the previously determined referral points and the individual was referred from the second stage to the third stage.

2.3.3. *Evaluation criteria*

The performance of the MST based on ability estimation was assessed according to mean RMSE and bias criteria that are frequently used in MST studies (Xiao & Bulut, 2022; Kim et al., 2015; Park, 2015; Sari & Raborn, 2018; Zheng, 2014). RMSE, and bias values for each simulation condition were calculated as follows:

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

$$Bias = \frac{\sum_{j=1}^{N}(\hat{\theta}_j - \theta_j)}{N},$$

where $\hat{\theta}_j$ and θ_j, j examinee predicted and true ability values; N is the total number of examinee. Multivariate analysis of variance (ANOVA) was used to test the significance of the MST variables in various conditions on the mean RMSE and bias values. 'Bonferroni' multiple comparison test was used to find out between which conditions the differences between the means were between. When interpreting the effect size, 0.00-0.19 was taken as very small, 0.20-0.49 as a small, 0.50-0.79 as medium, and 0.80 and larger as large effect sizes (Cohen, 1988).

3. FINDINGS

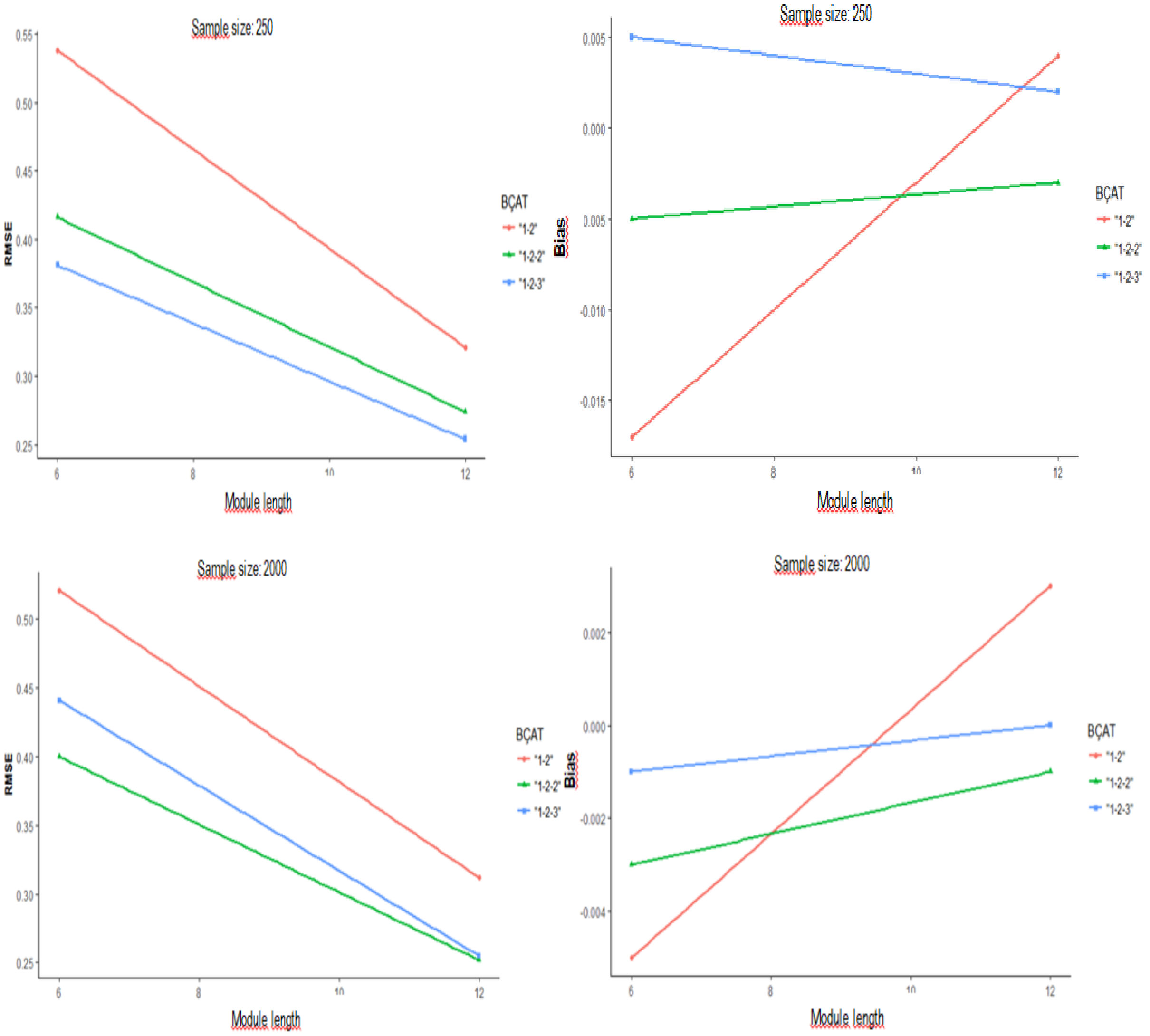
3.1. Findings on the Top-down Test Assembly Methods

We examined how the precision of ability estimation changes according to model lengths (6 and 12), panel patterns ("1-2", "1-2-2" and "1-2-3") and sample sizes (250 and 2000) in the top-down test assembly method in the MST application. In line with this, in order to interpret the findings, firstly MSTs were created in accordance with the simulation conditions in the problem. The findings related to ability estimation in MSTs created according to a variety of simulation conditions are presented in Table 1 and Figure 2.

Table 1. *Mean RMSE and bias values for MST created according to top-down test assembly method.*

Sample	Panel pattern	Module length	RMSE	Bias
250	“1-2”	6	0.538	-0.017
		12	0.321	0.004
	“1-2-2”	6	0.416	-0.005
		12	0.274	-0.003
	“1-2-3”	6	0.381	0.005
		12	0.254	0.002
2000	“1-2”	6	0.521	-0.005
		12	0.312	0.003
	“1-2-2”	6	0.400	-0.003
		12	0.252	-0.001
	“1-2-3”	6	0.441	-0.001
		12	0.255	0.000

Figure 2. *Plots of mean RMSE and bias values for MST created according to top-down test assembly method.*



As can be seen in Table 1, with the top-down test assembly method, the mean RMSE values obtained for different sample sizes, test lengths and panel patterns varied from 0.252 to 0.538. When the general lines of the results are investigated, the lowest error estimation was for the moderate length module applied to the large sample size with the '1-2-2' panel pattern, while the largest error estimation was for the short-length module applied to the small sample with the '1-2' panel pattern. When findings are investigated in terms of module length, for both sample sizes as module length increased, mean RMSE values appeared to reduce. When results are investigated in terms of panel pattern, the mean RMSE amount appeared to change for all test levels with the differentiation of panel patterns in small and large samples. In the transition from the '1-2' panel pattern to the '1-2-2' and '1-2-3' panel patterns, mean RMSE values fell. However, the mean RMSE values for both module lengths for '1-2-2' and '1-2-3' panel patterns applied to large samples were different with an increase for the transition from the '1-2-2' panel pattern to the '1-2-3' panel pattern. This increase was 0.041 for short module length and 0.003 for moderate module length in the transition from '1-2-2' panel pattern to the '1-2-3' panel pattern. When findings are investigated in terms of sample size, the increase in the sample size appeared to reduce mean RMSE values for both module lengths in all patterns, apart from the '1-2-3' panel pattern. For small samples, the lowest mean RMSE value was for the '1-2-3' panel pattern with a moderate length module, and for large samples, the lowest mean RMSE value was calculated for the '1-2-2' panel pattern with the moderate length module.

If the findings related to bias in Table 1 are investigated, mean bias values generally appear to be low. When the top-down test assembly method is chosen, the mean bias values vary from -0.017 to 0.005 for sample size, panel pattern and module length simulation conditions. The highest mean bias values were for '1-2' panel patterns applied to small samples with short module lengths. This was followed by short module length in small samples with '1-2-2' and '1-2-3' patterns, and the '1-2' panel pattern applied to large samples. The lowest mean bias value was calculated for the '1-2-3' panel pattern with moderate module length applied to large samples. This value was 0.000; in other words, this simulation condition had unbiased calculations. When findings were investigated in terms of module length, as module length increased, bias in panel patterns for both sample types was concluded to be reduced. When results are investigated in terms of panel pattern, for both module lengths, in small and large samples, the transition from the '1-2' panel pattern to '1-2-2' panel pattern and from '1-2-2' panel pattern to '1-2-3' panel pattern appeared to cause a fall in mean bias values. When findings are investigated in terms of sample size, as the sample size increased, the mean bias values were observed to fall by a small amount.

Within the scope of the subproblem in the research, whether the module length, panel pattern and sample size had statistically significant effects on the mean RMSE and bias findings obtained according to the top-down test assembly method was tested with the versatile ANOVA test. The F value and effect sizes (η^2) obtained from the ANOVA test are presented in Table 2.

Table 2. *Mean RMSE and ANOVA results for mean RMSE and bias values obtained when top-down test assembly method is chosen.*

	Evaluation Criteria					
		RMSE			Bias	
Study Conditions	*df*	*F*	η^2	*df*	*F*	η^2
Module length (M)	1	3379.332*	0.051	1	20.662*	0.049
Panel pattern (P)	2	404.320*	0.012	2	1.841	0.007
Sample (S)	1	0.034	0.052	1	3.753	0.007
P*M	2	50.648*	0.015	2	2.51	0.014
P*S	2	27.878*	0.008	2	8.395*	0.042
M*S	1	10.489*	0.001	1	0.686	0.014
P*M*S	2	12.019*	0.005	2	6.059*	0.028

*p<0.05

As observed in Table 2, the mean RMSE value obtained according to the top-down test assembly method significantly differed according to module length and panel pattern ($F_{1\text{-}358(module\ length)}$ = 3379.332, $p < 0.05$; $F_{2\text{-}357(Panel\ pattern)}$ = 404.320, $p < 0.05$). The eta-square values showed the efficacy of the module length and panel pattern on mean RMSE value was at moderate levels and the effect size was very small ($\eta^2_{(module\ length)}$ = 0.051, $\eta^2_{(Panel\ pattern)}$ = 0.012). To identify which panel patterns caused the difference among the panel patterns, the Bonferroni two-way comparison test was performed. According to the results of the test, the mean RMSE value was more affected by the '1-2-3' panel pattern ($\bar{X}$ = 0.423) compared to the '1-2-2' panel pattern ($\bar{X}$ = 0.335) and '1-2' panel pattern ($\bar{X}$ = 0.333). Additionally, the effects of the interactions of panel pattern-module length ($F_{4\text{-}355(P*M)}$ = 50.648, $p < 0.05$), panel pattern-sample size ($F_{4\text{-}355(P*S)}$ = 27.878, $p < 0.05$), module length-sample size ($F_{3\text{-}356(M*S)}$ = 10.489, $p < 0.05$) and panel pattern-module length-sample size ($F_{6\text{-}353(P*M*S)}$ = 12.019, $p < 0.05$) on mean RMSE values were significant. The panel pattern-module length ($\eta^2_{(P*M)}$ = 0.015), panel pattern-sample size ($\eta^2_{(P*S)}$ = 0.008), module length-sample size ($\eta^2_{(M*S)}$ = 0.001) and panel pattern-module length-sample size ($\eta^2_{(P*M*S)}$ = 0.005) had small levels of effect on mean RMSE value. However, the sample size did not significantly change the mean RMSE value.

As seen from Table 2, when the effects of module length, panel pattern and sample size on the mean bias values obtained according to the top-down test assembly method were examined, the mean bias values only appeared to significantly differ according to module length ($F_{1\text{-}358(module\ length)}$ = 20.662, $p < 0.05$). This finding is supported by the eta-square value ($\eta^2_{(module\ length)}$ = 0.049). Panel pattern and sample size did not cause a significant change in mean bias values. When significant effects of the interactions of these three variables on mean bias value were examined, panel pattern-sample size ($F_{4\text{-}355(P*S)}$ = 8.395, $p < 0.05$) and panel pattern-module length-sample size ($F_{6\text{-}353(P*M*S)}$ = 6.059, $p < 0.05$) interactions caused significant differences in mean bias value. Additionally, the effect of these variables on mean bias was at moderate levels ($\eta^2_{(P*S)}$ = 0.042, $\eta^2_{(P*M*S)}$ = 0.028).

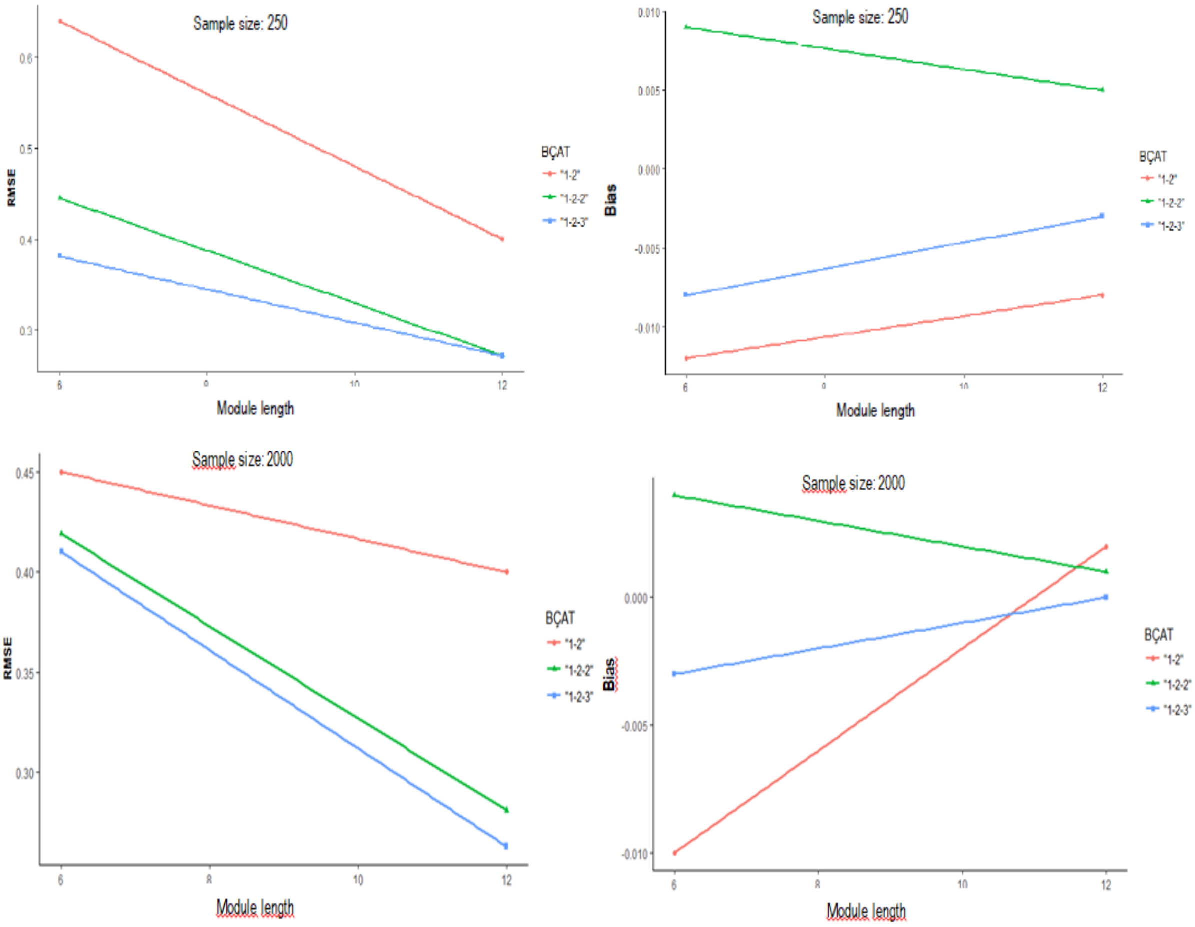
3.2. Findings on the Bottom-up Test Assembly Methods

Findings related to the change in the precision of ability estimations according to module lengths (6 and 12), panel patterns (1-2, 1-2-2 and 1-2-3) and sample sizes (250 and 2000) with the bottom-up test assembly method for MST applications are presented in Table 3 and Figure 3.

Table 3. *Mean RMSE and bias values for MST created According to bottom-up test assembly method.*

Sample	Panel pattern	Module length	RMSE	Bias
250	"1-2"	6	0.639	-0.012
		12	0.400	-0.008
	"1-2-2"	6	0.445	0.009
		12	0.272	0.005
	"1-2-3"	6	0.381	-0.008
		12	0.272	-0.003
2000	"1-2"	6	0.450	-0.010
		12	0.400	0.002
	"1-2-2"	6	0.419	0.004
		12	0.281	0.001
	"1-2-3"	6	0.410	-0.003
		12	0.263	0.000

Figure 3. *Plots of mean RMSE and bias values for MST created according to bottom-up test assembly method.*



As can be seen in Table 3, the mean RMSE values obtained for different module lengths, panel patterns and sample sizes according to the bottom-up test assembly method varied from 0.263 to 0.639. The lowest mean RMSEA estimation was for the '1-2-3' panel pattern with a moderate length module applied to a large sample, while the highest mean error estimation was for the '1-2' panel pattern with a small length module applied to a small sample. When the findings were investigated in terms of module length, for both sample sizes as the module length increased, the mean RMSE value appeared to reduce. When the findings were investigated according to panel pattern, for large and small samples, the differentiation of panel patterns changed the mean RMSE amount at all test levels. In the transition from the '1-2' panel pattern to '1-2-2' and '1-2-3' panel pattern, the mean RMSE values fell. When the findings were examined in terms of sample size, the increase in sample size appeared to reduce the mean RMSE values in many conditions. However, for the small sample, the moderate module length and '1-2-2' panel pattern, there was an increase of 0.09 when the same module length and panel pattern were applied to large samples. Additionally, when the '1-2-3' panel pattern was applied with short module length to small and large samples, a 0.29 increase was noticed. In small samples, the lowest mean RMSE value was calculated for the moderate length module with '1-2-2' and '1-2-3' panel patterns, while for large samples, the lowest mean RMSE value was obtained with the moderate length module applied in '1-2-3' panel pattern.

When the results relating to the bias obtained according to the bottom-up test assembly method are examined in Table 3, it appears that the mean values of the bias are generally very low. When the bottom-up test assembly method was chosen, the mean bias values according to module length, panel pattern and sample size simulation conditions varied from -0.012 to 0.009. The highest mean bias value belonged to the small sample with a short module length in the '1-

2' panel pattern. This pattern with short module length was followed by large samples with '1-2' patterns, then by short module length and small samples with '1-2-2' panel patterns. The lowest mean bias value for large samples was calculated for moderate module length with the '1-2-3' panel pattern. In these conditions, the calculated 0.000 mean bias value indicated bias-free calculations were performed. When the findings were examined in terms of module length, as module length increased, the bias for panel patterns in both sample types appeared to reduce. When the findings were examined in terms of panel patterns, for both module lengths with small and large samples, the transitions from the '1-2' panel pattern to '1-2-2' and from '1-2-2' panel pattern to '1-2-3' panel pattern reduced mean bias values. When the findings were examined in terms of sample size, as the sample size increased, the mean bias values appeared to reduce.

Within the scope of this subproblem, whether the effect of module size, panel pattern and sample size were statistically significant on mean RMSE and bias findings obtained according to the bottom-up test assembly method was tested with the versatile ANOVA test. The F value and effect sizes (η^2) obtained from the ANOVA test are presented in Table 4.

Table 4. *Mean RMSE and ANOVA results for mean RMSE and bias values obtained when bottom-up test assembly method is chosen.*

	Evaluation Criteria					
	RMSE			Bias		
Study Conditions	*df*	*F*	η^2	*df*	*F*	η^2
Module length (M)	1	2721.284*	0.032	1	6.400*	0.016
Panel pattern (P)	2	1000.355*	0.023	2	22.277*	0.105
Sample (S)	1	119.354*	0.001	1	1.786	0.005
P*M	2	10.654*	0.002	2	7.451*	0.033
P*S	2	140.741*	0.003	2	0.324	0.005
M*S	1	117.107*	0.001	1	0.592	0.005
P*M*S	2	149.044*	0.003	2	4.561*	0.022

*p<0.05

As can be seen in Table 4, the mean RMSE values obtained according to the bottom-up test assembly method differed significantly according to module length, panel pattern and sample size ($F_{1\text{-}358(module\ length)}$ = 2721.284, p < 0.05; $F_{2\text{-}357(Panel\ pattern)}$ = 1000.355, p < 0.05; $F_{1\text{-}358(sample)}$ = 119.354, p < 0.05). Module length and panel pattern had moderate effect on mean RMSE, while sample size had small effect ($\eta^2_{(module\ length)}$ = 0.032, $\eta^2_{(Panel\ pattern)}$ = 0.023, $\eta^2_{(sample)}$ = 0.001). To identify which panel patterns caused the difference, the Bonferroni two-way comparison test was performed. According to the test results, the '1-2-3' panel pattern ($\bar{X}$ = 0.472) had more effect on mean RMSE value compared to the '1-2-2' panel pattern ($\bar{X}$ = 0.356) and '1-2' panel pattern ($\bar{X}$ = 0.332). Additionally, the interactions of panel pattern-module length ($F_{4\text{-}355(P*M)}$ = 10.654, p < 0.05), panel pattern-sample ($F_{4\text{-}355(P*S)}$ = 140.741, p < 0.05), module length-sample ($F_{3\text{-}356(M*S)}$ = 117.107, p < 0.05) and panel pattern-module length-sample ($F_{6\text{-}353(P*M*S)}$ = 149.044, p < 0.05) had significant effects on mean RMSE value. The effect of these variables on mean RMSE was small ($\eta^2_{(P*M)}$ = 0.002, $\eta^2_{(P*S)}$ = 0.003, $\eta^2_{(M*S)}$ = 0.001, $\eta^2_{(P*M*S)}$ = 0.003).

As seen in Table 4, when the effects of module length, panel pattern and sample size on the mean bias values obtained according to the bottom-up test assembly method are examined, mean bias value differed significantly according to panel pattern and module length ($F_{1\text{-}358\ (module\ length)}$ = 6.400, p < 0.05; $F_{2\text{-}357(Panel\ pattern)}$ = 22.277, p < 0.05). The effect of module length on mean bias was small ($\eta^2_{(module\ length)}$ = 0.016), while the effect of panel pattern was at moderate levels ($\eta^2_{(Panel\ pattern)} = 0.105$). The Bonferroni two-way comparison test was performed to identify which panel pattern caused the difference. According to the test results, the '1-2-2' panel pattern ($\bar{X} = 0.008$) was more effective on mean RMSE value compared to the '1-2' panel pattern ($\bar{X} = 0.006$). However, sample size did not cause a significant difference in mean bias values. The interactions of panel pattern-module length and panel pattern-module length-sample size were observed to cause a significant difference in mean bias values ($F_{4\text{-}355(P{*}M)} = 7.451$, $p < 0.05$; $F_{6\text{-}353(P{*}M{*}S)} = 4.561$, $p < 0.05$). The effect of these variables on mean bias was at moderate levels ($\eta^2_{(P{*}M)} = 0.033$, $\eta^2_{(P{*}M{*}S)} = 0.022$).

4. DISCUSSION and CONCLUSION

Within the scope of the research, the performances of MSTs created according to top-down and bottom-up test assembly methods tested for module length, panel pattern and sample size using an item pool created from a real data set were compared. The MST components were module length, panel pattern and sample size. However, the study attempted to identify the correlation of these components with the test assembly method. For this reason, the focal point of the study was the top-down and bottom-up test assembly method recommended for combining MST panels introduced to the literature by Luecht and Nungester (1998). The research findings first showed that the module length affected mean RMSE and bias values with the top-down and bottom-up test assembly methods. For both test assembly methods, the moderate module length produced lower mean RMSE and bias values compared to the short module length. The probable reason for the difference in mean RMSE and bias values calculated for short and moderate module lengths may be the total item count. This situation may be interpreted as showing that as the total number of items in the test increases, the mean RMSE and bias values reduce. This finding is parallel to the findings of the study by Sari (2016) using the top-down test assembly method creating MST and CAT according to test management, content count and test length variables and comparing the performance of these two test types. In their study, they concluded that only test length had a significant effect on the mean RMSE value. This finding is also supported by the study of Yang (2016). In their study, the top-down test assembly method was used and as the test length increased, the RMSE and standard error values reduced. When the test length was 60, the bias was minimum, while it was maximum when the test length was 20. The mean bias values obtained according to the bottom-up test assembly method in this study significantly differed according to module length. For both samples, panel patterns with short module length had highest mean bias, while panel patterns with moderate module length had the smallest mean bias value. The study by Hembry (2014) studied the effect of two test lengths of short and moderate in MSTs created using the bottom-up test assembly method. This study had mean bias measures very close to zero and panel patterns with short test lengths had reduced mean RMSE and bias values. This finding is parallel to the findings in our research. Other similar findings were obtained in studies by Kim et al. (2013) using an OTB program as the test assembly method, Lynn Chen (2010) using the top-down test assembly method and Lu (2010) using the bottom-up test assembly method. A study by Zheng (2014) using the top-down test assembly method did not find a consistent difference between different module lengths.

However, in addition to the top-down and bottom-up test assembly methods, there are some MST studies, though few, using NAMSS, one of the automatic assembly methods. One of these studies by Dallas (2014) studied the directive and point effects of MSTs created by using 10 and 20 module lengths. The results of the study were similar to the results obtained for module lengths affecting MSTs investigated according to top-down and bottom-up test assembly methods completed in this study.

As supported by the studies mentioned above, the effect of module length on mean RMSE and bias values and the reason for the fall in mean RMSE and bias values as module length increases may be due to MSTs comprising short tests having lower measurement sensitivity. Longer tests ensure higher classification accuracy and consistency (Crocker & Algina, 1986; Luo, 2020).

Another finding in the research is the effect of panel patterns in top-down and bottom-up test assembly methods on mean RMSE and bias values. The change from the '1-2' panel pattern to '1-2-2' and '1-2-3' panel patterns according to the top-down test assembly method reduced mean RMSE and bias values in many conditions, while for the bottom-up test assembly method, it reduced mean RMSE and bias values in all conditions. This finding may be interpreted as showing that the increase in stage numbers in MST panels reduces the mean RMSE and bias. This finding is supported by the findings of a study comparing three-stage and two-stage MSTs by Patsula (1999), which found that three-stage MSTs produced less measurement error than two-stage MSTs. Additionally, the findings obtained from this study are consistent with the results of a study based on the 3 PL model by Zenisky (2004). Another similar finding was encountered in the study by Hembry (2014). In this study, MSTs created using the bottom-up test assembly method were investigated in four panel patterns of '1-3', '1-5'. '1-3-3' and '1-5-5'. Very small differences were obtained for estimated ability and mean bias values for the four panel patterns. Generally, mean bias measures very close to zero were obtained, as in this study. RMSE values were lower for the two-stage tests, different to the findings of this study. However, the difference between panel patterns was reported to be very low in this study. Additionally, there was a significant difference between '1-2', '1-2-2' and '1-2-3' panel patterns according to both methods in this research, with the '1-2-3' panel pattern concluded to have more effect on mean RMSE and bias values compared to other patterns. Sari (2016) obtained different findings in a study completed with the bottom-up test assembly method. In this study, the effects of the two-stage '1-3' and three-stage '1-3-3' panel patterns on RMSE were investigated, and it was reported that no significant difference was found. In research applying the bottom-up test assembly method, Yang (2016) obtained similar results to Sari (2016). In this study, four-panel structures were investigated ("1-3", "1-5", "1-3-3" and "1-5-5") and significant differences were not found. Another parallel finding to these studies was obtained in the study by Jodoin et al. (2006) and Luo and Kim (2018). Studies by Zheng et al. (2012) and Zheng (2014) used the top-down test assembly method and reported no significant differences were found between four-stage models and three-stage models. As can be seen, there are two different results about the effect of panel patterns on MST studies. The probable cause for the different results may be other variables that were fixed in both studies. In fact, it should not be ignored that increasing the number of stages in the panel structure may provide better measurement sensitivity as it is directly proportional to the individual's responses to higher numbers of items.

According to the research findings, for the top-down test assembly method, for '1-2' and '1-2-2' panel patterns with short and moderate module length, the increase in sample size lowered the mean RMSE and bias values. For the bottom-up test assembly methods, the increase in sample size with the '1-2' panel pattern for short and moderate module lengths, the '1-2-2' panel pattern with short module length and the '1-2-3' panel pattern with moderate module length lowered mean RMSE and bias values. Additionally, for the bottom-up test assembly method, the sample size had a statistically significant effect on mean RMSE and bias values, while for the top-down test assembly method, it was concluded there was no significant effect. In this context, no definite conclusion can be made about which test assembly method should be chosen for small or large sample sizes. In fact, in international studies of MST, the use of large-scale tests is an indicator of applicability for large samples. Based on the research findings, interpretations can be made about the applicability of the bottom-up test assembly method using '1-2' panel pattern with short and moderate module lengths, '1-2-2' panel pattern for short module lengths and '1-2-3' panel pattern with moderate module length for large samples. The reason for choosing sample size as a variable in the research is to ensure the ability to see possible outcomes when MST's, used for samples in large-scale international tests, are applied to institutional exams like for inspectors, specialists, and judges, completed with smaller samples in our country, or even in lesson selection exams applied in middle schools and high schools, in the future if appropriate computer infrastructure is developed. In a similar study investigating small sample sizes, Yan et al. (2014) investigated MSTs in accordance with the 'tree-based' approach. The study concluded that MSTs applied to small samples displayed good performance.

When assessed as a whole, the top-down and bottom-up test assembly methods produced similar findings and both methods are recommended for use in creating MSTs. Additionally, the research investigated the top-down and bottom-up test assembly methods among automatic test assembly methods. Later studies are recommended to study other methods like ASM, NAMSS and maximum priority index, in addition to these methods, linear programming methods and the test assembly method performed at the time of the exam called the 'on-the-fly' test assembly method in the literature. In the research, item and ability parameters suitable for only the 2 PL model were estimated as the item pool was created according to a real test set and the MST was created accordingly. In later studies, parameters may be estimated according to 2 PL and 3 PL models to research the effect of logistic models on MST performance.

Acknowledgments

This paper was produced from the first author's doctoral dissertation prepared under the supervision of the second author.

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 authors.

Contribution of Authors

Ebru Doğruöz: Investigation, Resources, Visualization, Software, Analysis, and Writing-original draft. **Hülya Kelecioğlu:** Methodology, Supervision, and Critical Review.

Orcid

Ebru Doğruöz https://orcid.org/0000-0001-6572-274X
Hülya Kelecioğlu https://orcid.org/0000-0002-0741-9934

REFERENCES

American Institute of Certified Public Accountants. (2019, February 18). *CPA exam structure*. https://www.aicpa.org/becomeacpa/cpaexam/examinationcontent.html

Belov, D.I. (2016). *Review of modern methods for automated test assembly and item pool analysis*. Law School Admission Council Research Report 16-01 March 2016, LSAC Research Report Series, 23 pages, https://www.lsac.org/docs/default-source/research-(lsac-resources)/rr-16-01.pdf

Breithaupt, K., Ariel, A., & Veldkamp, B. (2005). Automated simultaneous assembly for multistage testing. *International Journal of Testing*, *5(*3), 319-330. https://doi.org/10.1207/s15327574ijt05038

Breithaupt, K., & Hare, D.R. (2007). Automated simultaneous assembly of multistage testlets for a high-stakes licensing examination. *Educational and Psychological Measurement*, *67*(1), 5-20. https://doi.org/10.1177/0013164406288162

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. CBS College Publishing.

Dallas, A. (2014). *The effects of routing and scoring within a computer adaptive multi-stage framework* [Unpublished Doctoral Dissertation]. The University of North Carolina.

Davis, L.L., & Dodd, B.G. (2003). Item exposure constraints for testlets in the verbal reasoning section of the MCAT. *Applied Psychological Measurement*, *27*(5), 335-356. https://doi.org/10.1177/0146621603256804

Educational Testing Service. (2018, February 18). *Computer-delivered GRE general test content and structure.* http://www.ets.org/gre/revised%5Cgeneral/about/content/computer/

Hambleton, R.K., & Xing, D. (2006). Optimal and nonoptimal computer-based test designs for making pass-fail decisions. *Applied Measurement in Education, 19*(3), 221-239. https://doi.org/10.1207/s15324818ame1903_4

Hembry, I.F. (2014). *Operational characteristics of mixed format multistage tests using the 3PL testlet response theory model* [Unpublished Doctoral Dissertation]. University of Texas at Austin.

Hendrickson, A. (2007). An NCME instructional module on multistage testing. *Educational Measurement: Issues and Practice, 26*(2), 44-52. https://doi.org/10.1111/j.1745-3992.2007.00093.x

Hogan, J., Thornton, N., Diaz-Hoffmann, L., Mohadjer, L., Krenzke, T., Li, J. & Khorramdel, L. (2016, July 5,). US program for the international assessment of adult competencies (PIAAC) 2012/2014: Main study and national supplement technical report (NCES 2016-036REV). U.S. Department of Education. National Center for Education Statistics. https://nces.ed.gov/pubs2016/2016036 rev.pdf

Jodoin, M.G., Zenisky, A., & Hambleton, R.K. (2006). Comparison of the psychometric properties of several computer-based test designs for credentialing exams with multiple purposes. *Applied Measurement in Education, 19*(3), 203-220. http://doi.org/10.1207/s15324818ame1903_3

Khorramdel, L., Pokropek, A., & van Rijn, P. (2020). Special Topic: Establishing comparability and measurement invariance in large-scale assessments, part I. *Psychological Test and Assessment Modeling*, *62*(1), 3-10. https://www.psychologie-aktuell.com/fileadmin/Redaktion/Journale/ptam-2020-1/01_Khorramdel.pdf

Kim, S., Moses, T., & You, H. (2015). A comparison of IRT proficiency estimation methods under adaptive multistage testing. *Journal of Educational Measurement*, *52*(1), 70-79. https://doi.org/10.1111/jedm.12063

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

Kirsch, I., & Lennon, M.L. (2017). PIAAC: A new design for a new era. *Large-scale Assessments in Education*, *5*, 11. https://doi.org/10.1186/s40536-017-0046-6

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

Luecht, R. (2000). *Implementing the 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 (NCME), New Orleans, LA.

Luecht, R.M. (2006). *Designing tests for pass-fail decisions using item response theory.* In S. Downing & T. Haladyna (Eds.), Handbook of test development, 575-596. Lawrence Erlbaum Associates.

Luecht, R., Brumfield, T., & Breithaupt, 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. https://www.learntechlib.org/p/87698/.

Luo, X. (2019). Automated test assembly with mixed-ınteger programming: The effects of modeling approaches and solvers. *Journal of Educational Measurement*, *57*(4), 547-565. https://doi.org/10.1111/jedm.12262

Luo, X., & Kim, D. (2018). A top-down approach to designing the computerized adaptive multistage test. *Journal of Educational Measurement*, *55*(2), 243-263. https://doi.org/10.1111/jedm.12174

Lynn 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.

OECD (2015). *PISA 2015 technical report.* http://www.oecd.org/pisa/sitedocument/PISA-2015-Technical-Report-Chapter-1-Programme-for-International-Student-Assessment-an-Overview.pdf

OECD (2017). *PISA 2015 technical report.* http://www.oecd.org/pisa/sitedocument/PISA-2015-Technical-Report-Chapter-9-Scaling-PISA-Data.pdf

Papadimitriou, C.H., & Steiglitz, K. (1982). *Combinatorial optimization: Algorithms and complexity*. Prentice-Hall.

Park R. (2015). *Investigating the impact of a mixed-format item pool on optimal test designs for multistage testing* [Unpublished doctoral dissertation]. University of Texas, Austin.

Patsula, L.N. (1999). *A comparison of computerized-adaptive testing and multi-stage testing* [Unpublished doctoral dissertation]. University of Massachusetts at Amherst.

Pihlainen, K.A.I., Santtila, M., Häkkinen, K., & Kyröläinen, H. (2018). Associations of physical fitness and body composition characteristics with simulated military task performance. *The Journal of Strength & Conditioning Research*, *32*(4), 1089-1098. https://doi.org/10.1519/jsc.0000000000001921

Sari, H.İ. (2016). *Examining content control in adaptive tests: Computerized adaptive testing vs. computerized multistage testing* [Unpublished doctoral dissertation]. University of Florida.

Sari, H.I., & Raborn, A. (2018). What information works best? A comparison of routing methods. *Applied psychological measurement*, *42*(6), 499-515. https://doi.org/10.1177/0146621617752990

Şahin Kürşad, M., Çokluk-bökeoglu, Ö. & Çıkrıkçı, N. (2022). The study of the effect of item parameter drift on ability estimation obtained from adaptive testing under different conditions. *International Journal of Assessment Tools in Education*, *9*(3), 654-681. https://doi.org/10.21449/ijate.1070848

Theunissen, T.J.J.M. (1985). Binary programming and test design. *Psychometrika*, *50*(4), 411-420. https://link.springer.com/article/10.1007/BF02296260

Tian, C. (2018). Comparison of four stopping rules in computerized adaptive testing and examination of their application to on-the-fly multistage testing [Unpublished master dissertation]. University of Illinois.

Van der Linden, W.J., & Glas, C.A.W. (2000). Capitalization on item calibration error in adaptive testing. *Applied Measurement in Education*, *13*, 35-53. https://doi.org/10.1207/s15324818ame1301_2

van der Linden, W.J. (2005). *Linear models of optimal test design.* Springer.

van der Linden, W.J., & Boekkooi-Timminga, E. (1989). A maximin model for IRT-based test design with practical constraints. *Psychometrika*, *54*(2), 237-247. https://link.springer.com/article/10.1007/BF02294518

Veldkamp, B.P. (1999). Multiple-objective test assembly problems. *Journal of Educational Measurement, 36*, 253-66. http://www.jstor.org/stable/1435157

Veldkamp, B.P., Matteucci, M., & de Jong, M.G. (2013). Uncertainties in the item parameter estimates and robust automated test assembly. *Applied Psychological Measurement*, *37*, 123-139. https://doi.org/10.1177/0146621612469825

Wang, K. (2017). *Fair comparison of the performance of computerized adaptive testing and multistage adaptive testing* [Unpublished doctoral dissertation]. Michigan State University.

Wise, S.L., & Kingsbury, G.G. (2000). Practical issues in developing and maintaining a computerized adaptive testing program. *Psicologica*, *21*, 135-155. https://www.uv.es/revispsi/articulos1y2.00/wise.pdf

Xiao, J., & Bulut, O. (2022). Item selection with collaborative filtering in on-the-fly multistage adaptive testing. *Applied Psychological Measurement*, *46*(8), 690-704. https://doi.org/10.1177/01466216221124089

Xing, D., & Hambleton, R.K. (2004). Impact of test design, item quality, and item bank size on the psychometric properties of computer-based credentialing examinations. *Educational and Psychological Measurement*, *64*, 5-21. https://doi.org/10.1177/0013164403258393

Xu, L., Wang, S., Cai, Y., & Tu, D. (2021). The automated test assembly and routing rule for multistage adaptive testing with multidimensional item response theory. *Journal of Educational Measurement*, *58*, 538-563. https://doi.org/10.1111/jedm.12305

Yan, D., Lewis, C., & von Davier, A. (2014). Overview of computerized multistage tests. In D. Yan, A.A. von Davier, & C. Lewis (Eds.). *Computerized Multistage Testing: Theory and Applications*, 3-20. Chapman & Hall.

Yan, D., von Davier, A.A., & Lewis, C. (Eds.). (2014). *Computerized Multistage Testing: Theory and Applications (1st ed.)*. Chapman and Hall/CRC. https://doi.org/10.1201/b16858

Yang, L. (2016). *Enhancing item pool utilization when designing multistage computerized adaptive tests* [Unpublished doctoral dissertation]. Michigan State University.

Zenisky, A. (2004). *Evaluating the effects of several multistage testing design variables on selected psychometric outcomes for certification and licensure assessment* [Unpublished doctoral dissertation]. University of Massachusetts at Amherst.

Zenisky, A., & Hambleton, R. (2014). Multistage test designs: Moving research results into practice. In Yan, D., Von Davier, A., & Lewis, C. (Eds.), *Computerized Multistage Testing: Theory and Applications*, 21-36. Chapman & Hall.

Zenisky, A., Hambleton, R.K. & Luecht, R.M. (2010). Multistage testing: Issues, designs and research. In: der Linden, W.J. & Glas, C.A.W. (Eds.). *Elements of Adaptive Testing*. 355-372. Springer.

Zenisky, A.L., Sireci, S.G., Martone, A., Baldwin, P., & Lam, W. (2009). Massachusetts adult proficiency tests technical manual supplement: 2008-2009. *Center for Educational Assessment Research*. http://www.umass.edu/remp/docs/MAPTTMSupp7-09 final.pdf

Zheng, Y. (2014). *New methods of online calibration for item bank replenishment* [Unpublished Doctoral Dissertation]. University of Illinois at Urbana-Champaign.

Zheng, Y., & Chang, H.-H. (2015). On-the-Fly assembled multistage adaptive testing. *Applied Psychological Measurement*, *39*(2), 104-118. https://doi.org/10.1177/0146621614544519

Zheng, Y., Nozawa, Y., Gao, X., & Chang, H. (2012). *Multistage adaptive testing for a large-scale classification test: the designs, heuristic assembly, and comparison with other testing modes*. Paper presented at the Annual Meeting of the National Council on Measurement in Education (NCME) (ACT Research Reports 2012-6). Vancouver, British Columbia, Canada.

Zheng, Y., Nozawa, Y., Zhu, R., & Gao, X. (2016). Automated top-down heuristic assembly of a classification multistage test. *Int. J. Quantitative Research in Education*, *3*(4), 242-265. https://doi.org/10.1504/IJQRE.2016.082387

Zimowski, M.F., Muraki, E., Mislevy, R.J., & Bock, R.D. (1996). *BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items*. [Computer software]. Scientific Software International.