International Journal of Assessment Tools in Education
2025, Vol. 12, No. 2, 414–429
https://doi.org/10.21449/ijate.1474855
journal homepage: https://dergipark.org.tr/en/pub/ijate *Research Article*

The effect of polytomous item ratio on ability estimation in multistage tests

Hasibe Yahsi Sari[1*], **Hulya Kelecioglu**[2]

[1]Necmettin Erbakan University, Eregli Faculty of Education, Department of Educational Sciences, Konya, Türkiye
[2]Hacettepe University, Faculty of Education, Department of Educational Sciences, Ankara, Türkiye

ARTICLE HISTORY
Received: Apr. 29, 2024
Accepted: Mar. 6, 2025

Keywords:
Multistage testing,
Polytomous item,
Mixed test,
Ability estimation,
Pisa 2018.

Abstract: The aim of the study is to examine the effect of polytomous item ratio on ability estimation in different conditions in multistage tests (MST) using mixed tests. The study is simulation-based research. In the PISA 2018 application, the ability parameters of the individuals and the item pool were created by using the item parameters estimated from the dichotomous and polytomous items obtained in the field of reading skills. MST conditions; panel design, test lengths, routing methods, and polytomous item ratio. Simulation data, MST pattern and analysis were obtained with the help of WinGen, CPLEX, and the "mstR" package in the R Studio program. A total of 108 conditions and 100 replications were examined in the study. As a result of the simulations, RMSE, mean absolute bias and correlation values were calculated. As a result of the research, it is seen that when the ratio of polytomous items in the tests increases from 10% to 50%, the mean absolute bias and RMSE values decrease while the correlation values increase. As the test length increases, RMSE and mean absolute bias values decrease while correlation values increase. In terms of routing methods, MFI performed better than the NC routing method. In general, three-stage panel designs gave significantly better results than two-stage panel designs. In 1-2 and 1-4 panel designs, it does not matter which routing method is used.

1. INTRODUCTION

Computer-based measurement and assessment applications in education have been utilized since the 1980s (Weiss, 1982; Weiss & Kingsbury, 1984). Over time, advancements in technology and the integration of artificial intelligence have significantly increased the importance of technology in educational measurement and assessment. Among these innovations, computerized adaptive testing (CAT) and multistage testing (MST) have gained global prominence. The first large-scale application of CAT was conducted in the United States with the Graduate Record Examination (GRE) in 1993. Some of the exams that have been computerized adaptive include The Law School Admission Test (LSAT), the Test of English as a Foreign Language (TOEFL), the National Council of Architectural Registration Boards (NCARB), the National Assessment of Educational Progress (NAEP), and the U.S. Medical Licensure Examination (USMLE) (Hendrickson, 2007). In recent years, CAT has been replaced by MST in national and international large-scale exams. GRE was implemented as MST in 2011, the Programme for International Student Assessment (PISA) in 2018, and the Programme for the International

*CONTACT: Hasibe YAHSİ SARİ ✉ hsbyahsi@gmail.com Necmettin Erbakan University, Eregli Faculty of Education, Department of Educational Sciences, Konya, Türkiye

e-ISSN: 2148-7456

Assessment of Adult Competencies (PIAAC) in 2012 (Yamamoto *et al*., 2019). Notably, in 2023, the Scholastic Aptitude Test (SAT), a standardized large-scale international test taken by students who want to attend university in the United States, was administered in digital format as MST. Finally, in 2024, the SAT administered throughout the U.S. had fully transitioned to MST.

The most significant advantage provided by CAT is the establishment of a relationship between the individual's ability and item difficulty. Unlike fixed test lengths in traditional paper-and-pencil test applications, personalized test lengths are available in CAT applications (Wang, 2017; Weiss, 1983). However, criticisms of CAT include varying test lengths, the inability to revisit previous questions, and differences in questions presented to test-takers. MST emerged as a solution to address some limitations of CAT while maintaining its adaptive nature. MST organizes tests into modules, stages, and panels. Test-takers are directed to subsequent modules based on their performance, creating a dynamic yet structured pathway for assessment (Zheng *et al*., 2012). MST applications increase in popularity due to their ability to tailor to individuals, allow test developers to preview test forms in advance, and enable individuals taking the test to review their answers. In MST, control over content and other features provides increased test security and content management (Hendrickson, 2007). While CAT adapts to individuals at the item level, MST adapts at the module level. MST encourage individuals with lower ability levels while preventing boredom in individuals with higher ability levels. These features have made MST increasingly popular for large-scale assessments, particularly those requiring precise and efficient measurement across broad ability ranges.

Tests composed of both dichotomously and polytomously scored items are called mixed-format tests. Mixed-format tests are crucial both due to their increasing prevalence in international large-scale tests and because they offer higher levels of test information compared to dichotomous items (Rosa *et al*., 2001). In recent years, many large-scale assessments, such as PIAAC and PISA, now employ mixed-format tests combining dichotomous (e.g., multiple-choice) and polytomous (e.g., open-ended) items (Yamamoto *et al*., 2019). In addition to mixed-format tests, the proportions of polytomous items used also play a significant role in the accuracy of ability estimation (Kim & Dodd, 2014). Kim *et al*. (2013) examined various routing methods and panel designs within the context of classification testing in MST applications based on the partial credit model. Their findings indicated that longer tests yielded higher accuracy, whereas a 50% pass rate resulted in the lowest accuracy levels. Park *et al*. (2014) explored a new item pool usage method for mixed-format tests in MST. Three designs were implemented using the linear programming (LP) test assembly method, with item change rates of 0.22, 0.44, and 0.66 for each test combination. The findings demonstrated that the applied MST recombination conditions enhanced item pool utilization while maintaining the desired MST structure.

1.1. Purpose and Importance of Research

The aim of this study is to examine the impact of varying polytomous item ratios in individually adapted multistage tests on the accuracy of ability estimation under different testing conditions. A review of the literature reveals numerous studies on tests comprising both dichotomous and polytomous items, where MST is analyzed under various conditions, including test assembly methods, routing strategies, and test lengths (Dogruoz, 2018; Kim *et al*., 2010; Luecht & Nungester, 1998; Luo & Kim, 2018; Yahsi Sari & Kelecioglu, 2023; Wang, 2017). While the importance of mixed-format tests is well-recognized, research on their application within MST frameworks remains limited (Kim *et al*., 2012; Kim *et al*., 2013). This study seeks to bridge this gap by investigating the interaction between polytomous item ratios and other critical testing variables, thereby making a substantial contribution to the literature.

The item ratio in MST with mixed-format tests is anticipated to exhibit significant variability under different conditions, highlighting the theoretical and practical importance of this study. In the international context, simulation data based on the item parameters of dichotomous and

polytomous items from real-world MST applications in the reading domain of PISA 2018 have been utilized, ensuring ecological validity. This approach offers valuable insights for future applications of PISA and other large-scale assessments. The statistical properties of the routing module are especially crucial, as they significantly affect the overall measurement accuracy of the test (Kim & Plake, 1993). In the national context, there is a clear gap in research addressing mixed-format tests within MST applications. While real-world CAT studies have been conducted at the research level (Cikrikci *et al*., 2020; Senel & Kutlu, 2018), studies focusing on MST applications are still emerging. This research will not only contribute to the international literature but also advance MST applications within the national context. Furthermore, there is limited research on the performance of test assembly methods under varying conditions with mixed-format tests, particularly those based on the linear programming approach-one of the automatic test assembly methods (Park *et al*., 2014; Park, 2015). To address these gaps, future studies should explore the application of linear programming test assembly methods to mixed-format tests, examining the effects of different panel designs, routing methods, and proportions of polytomous items. Such research would contribute significantly to advancing both the theory and practice of MST, supporting its widespread adoption in educational assessment. This study seeks to address the following research questions:

1- Primary Research Question:

- How does the ratio of polytomous items in a test (10%, 30%, and 50%) affect individuals' ability estimations?

2- Secondary Research Questions:

- How do test length (20, 40, and 60 items) and panel design (“1-2”, “1-3”, “1-4”, “1-2-2”, “1-2-3”, and “1-3-3”) influence the relationship between the ratio of polytomous items and ability estimations?
- How do routing methods (Maximum Fisher Information vs. Number-Correct) interact with the ratio of polytomous items in affecting ability estimations?

By exploring these questions, the study aims to provide practical recommendations for optimizing MST design and improving the precision of ability estimates across diverse testing conditions.

2. METHOD

2.1. Research Method

The aim of the research is to examine the effect of polytomous item ratios in individually adapted multistage mixed tests on ability estimation under different conditions. Systematic research aimed at generating new products or new processes, or significantly improving existing ones, by utilizing existing knowledge gained from research or experience is experimental development research (OECD, 2002). In the research, new conditions were determined that will improve the systems produced by simulation. In this aspect, the research is experimental development research based on simulation.

2.2. Sample of Research

In PISA 2018, a multistage computer-based test was administered in the domain of reading skills. For this research, data were generated through simulation based on the parameters of dichotomous and polytomous items from individuals who participated in the PISA 2018 assessment in the reading skills domain. According to the *PISA 2018 Technical Report Final Annex A*, 244 items-72 trend items and 172 new items-were utilized in the computerized multistage reading skills test. The data were obtained from the website of the Organization for Economic Co-operation and Development (OECD).

2.3. Research Design

In this section, the manipulated and fixed conditions in the simulation study are explained. The fixed conditions in the study include the sample size (10,000 individuals), the distribution of

individuals' ability levels (normal distribution, *N* (0,1)), and the method used for ability estimation for the ability levels calculated at the end of MST (Expected a Priori-EAP). EAP (Bock & Mislevy, 1982) is one of the popular methods that belong to the Bayesian ability estimations. There are several studies in the literature that use normal ability distribution and the EAP ability estimation method (Park, 2015; Sahin & Ozturk, 2019). Therefore, the EAP ability estimation method was used in this study. The manipulated conditions in the study are panel design, total test length, routing method, and the ratio of polytomous items in the total test length. These conditions were determined considering the most frequently used conditions in both simulation studies and real MST applications according to the existing literature.

In the panel design, six different panel designs were used: "1-2", "1-3", "1-4", "1-2-2", "1-2-3", and "1-3-3", based on Patsula (1999) "1-3" and "1-3-3", Zenisky (2004) "1–3–3", "1–2–3", "1–3–2", and Sarı and Raborn (2018) "1-3", "1-2-2", "1-2-3", and "1-3-3". For test lengths, Wang (2017) 45 and 60; Sarı and Raborn (2018) used test lengths of 30 and 60. Based on the existing literature, three different test lengths were determined as 20, 40 and 60, representing short, medium and long test lengths.

When we examine the literature in terms of routing method: Approximate Maximum Information (AMI), Defined Population Intervals (DPI), Modified Approximate Maximum Information (M-AMI), Stage-Level DPI (SL-DPI), and Module-Level DPI (ML-DPI) method are frequently used routing methods in the literature (Kim *et al*., 2010; Kim *et al*., 2013; Wang, 2017; Zenisky, 2004). Also, there are some studies that provide the maximum amount of information based on Maximum Likelihood Estimation (MLE) and M-AMI of ability (Kim *et al*., 2012). Number of Correct Responses (NC) is another common routing method in MST (Weissman *et al*., 2007; Zenisky *et al*., 2010). In recent years, using the mstR (Magis *et al*., 2018) package, Maximum Fisher Information (MFI), Maximum Likelihood Weighted Information (MLWI), Maximum Posterior Weighted Information (MPWI), Kullback–Leibler (KL), and Posterior Kullback–Leibler (KLP) routing methods have also begun to be used (Boztunc-Ozturk, 2019; Sarı & Raborn, 2018). In this study, as routing methods, two directing methods were determined: MFI and NC.

Park (2015) determined the polytomous item rates as (10%, 30%, 50% and 70%) in his study. In this study, the ratio of polytomous items to the total test length was determined as 10%, 30% and 50%. All conditions were crossed with each other, resulting in 108 (6x3x2x3) different conditions examined in the study. The simulation was repeated 100 times for all conditions. The manipulated conditions of the research are summarized in Table 1.

Table 1. *Research condition.*

Condition	Number of Levels	Lower Levels
Panel Design	6	1-2 1-3 1-4 1-2-2 1-2-3 1-3-3
Test Length	3	20 items 40 items 60 items
Routing Methods	2	MFI NC
Ratio of Polytomous Items	3	%10 %30 %50
Total	6x3x2x3=108	

2.4. Data Collection Process

In this stage, 200 polytomous and 400 dichotomous item parameters and ability levels (theta-θ) of 10,000 individuals were generated using the WinGen3 program (Han, 2007). The three stages are explained in detail.

2.4.1. *1. Stage: Generation of ability levels*

In the first stage, ability levels for a sample of 10,000 individuals were generated using the WinGen3 program according to a normal distribution (N (0,1)) with a mean of 0 and a standard deviation of 1. 100 iterations were performed to generate ability levels. The probability of responding to each item for individuals generated in the simulation was calculated according to the Generalized Partial Credit Model (GPCM).

2.4.2. *2. Stage: Obtaining data and conducting item analyses*

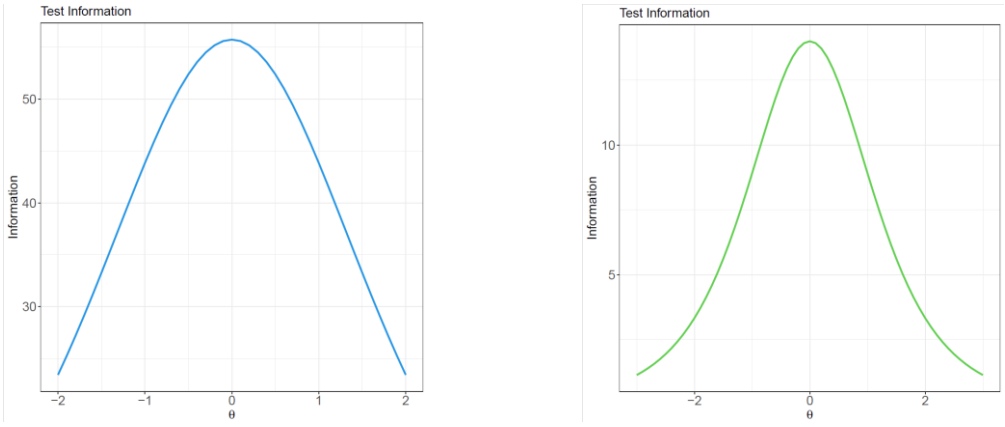
A simulation study based on the characteristics of item parameters used in the MST application in reading skills in PISA 2018 was conducted in the research. Among these items, 223 were dichotomous and 21 were polytomous. In this context, MST was also applied in the research, and parameters of both dichotomous and polytomous items were used. Descriptive statistics of item parameters were calculated using the R program (R Development Core Team, 2018). Descriptive statistics of the original items are provided in Table 2.

Table 2. *Descriptive statistics of item parameters.*

	2 categories		3 categories		
	a	b	a	b1	b2
Min.	0.15	-1.91	0.40	-0.40	-2.02
Max.	1.83	2.66	1.25	2.07	2.02
Mean	1.01	0.03	0.72	0.65	0.20

The test information functions of the item pools consisting of items used in the MST application in the reading skills domain of PISA 2018 are shown in Figure 1. When comparing the test information functions in Figure 1, we observe that polytomous items provide more information, despite being fewer in number than dichotomous items.

Figure 1. *Test information functions for PISA data.*



Dichotomous items Polytomous items

2.4.3. *3. Stage: Generation of Item Pool*

In this stage, item pools were generated using the estimated item parameters. Item parameters for dichotomous items were generated using the Two-Parameter Logistic (2PL) IRT model, while item parameters for polytomous items were generated using the GPCM. When generating

item parameters for the items in the item pool, the distribution of item parameters used in the reading skills domain of PISA 2018 was considered.

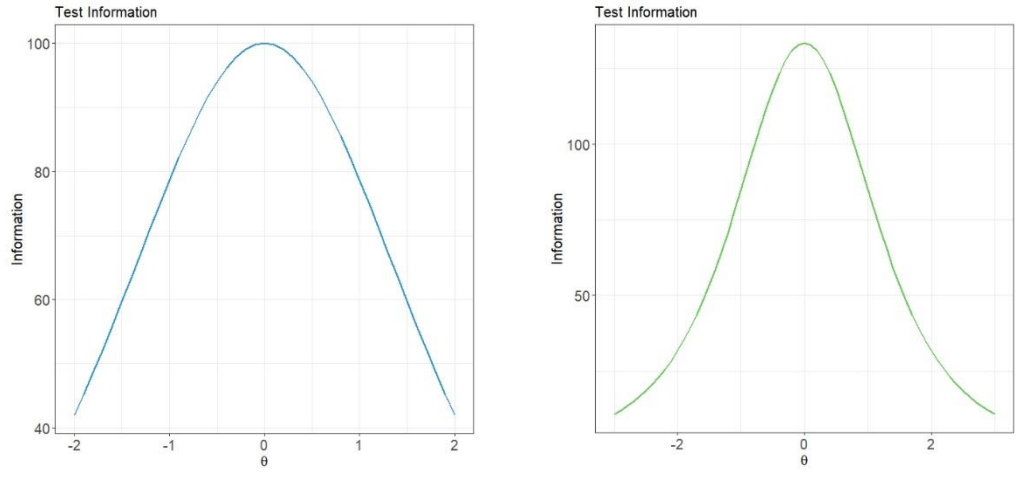
In MST studies conducted in the literature, the item pool size generally varies between 200 and 600 items (Lim, 2019; Xing & Hambleton, 2004; Zheng *et al*., 2012). In MST studies using mixed-item formats, 424 items have been used (Kim *et al*., 2012; Park *et al*., 2014). In the study by Kim *et al*. (2012), out of the 424-item pool, 244 items (57.55%) were dichotomous, 113 items (26.65%) were three-category, and 67 items (15.80%) were four-category items. In this study, a total of 600 items were generated for the item pool, with 400 being two-category and 200 being three-category items. During the panel construction stage, polytomous items were distributed into modules representing 10%, 30%, and 50% of the entire test. Descriptive statistics of the generated 400 and 200-item pools are provided in Table 3.

Table 3. *Descriptive statistics of item parameters in the 400 and 200 items pool.*

	400 items		200 items		
	a	b	a	b1	b2
Min.	0.15	-1.90	0.40	-0.40	-0.34
Max.	1.82	2.65	1.25	1.98	2.07
Mean	0.97	0.44	0.85	0.42	1.21

When examining the descriptive statistics of the PISA items and the simulated item parameters, slight differences were observed, which are likely due to the size of the item pool. The test information functions of the generated item pools consisting of 400 two-category and 200 three-category items are shown in Figure 2. When we examine the distribution of test information functions of PISA and simulation data, it is observed that the amount of test information obtained from the simulation is higher due to the larger number of items.

Figure 2. *Test information functions for item pools.*



Dichotomous items Polytomous items

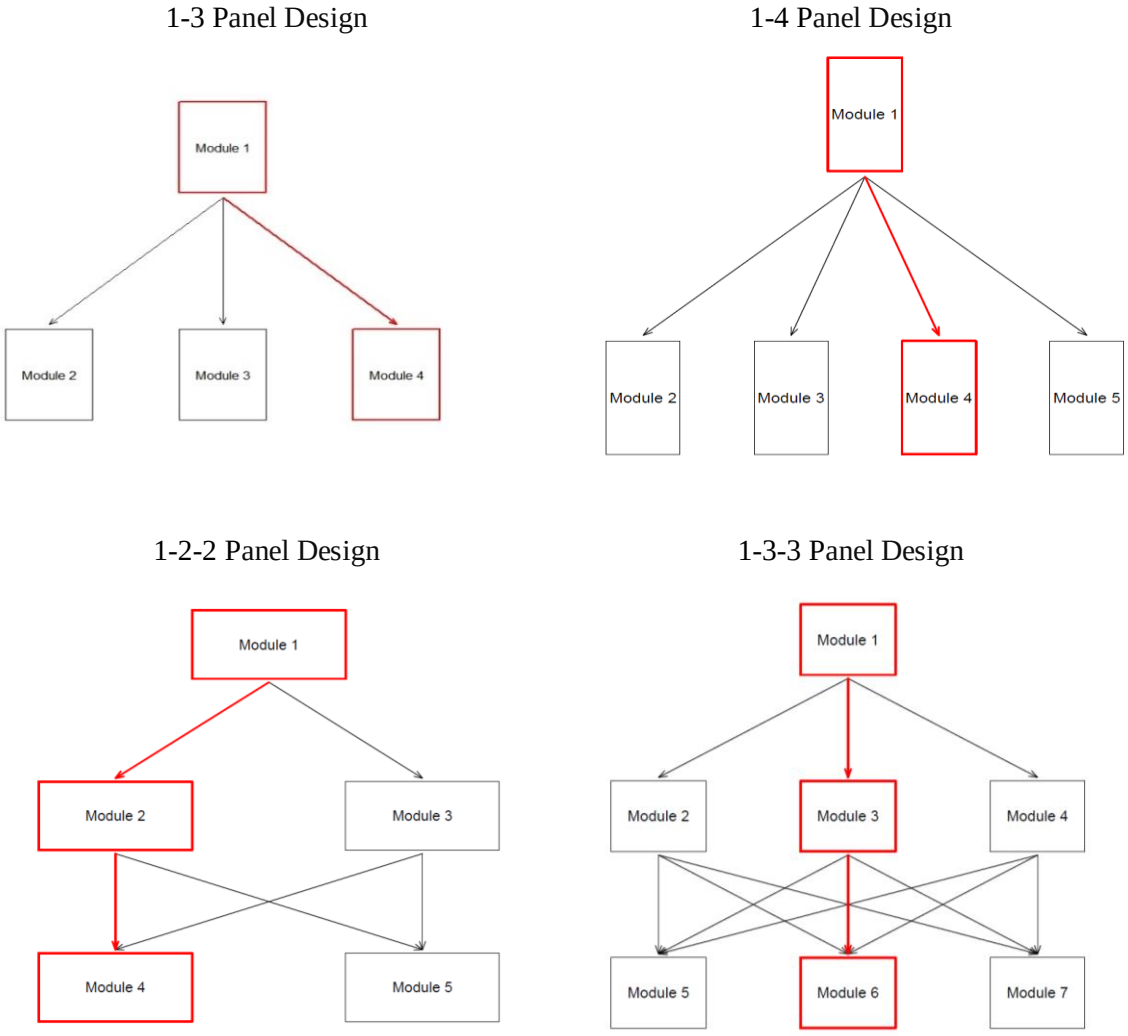
2.5. Data Analysis

An MST simulation was created as a data analysis tool. The MST simulation consists of modules and panels. There is a total of six-panel designs: 1-2, 1-3, 1-4, 1-2-2, 1-2-3, and 1-3-3. There are three test lengths: 20, 40, and 60. The number of items in the modules varies according to the total test length and panel designs.

2.5.1. *MST Simulation*

In the process of selecting items for panels in different designs in the MST, the following steps were taken. In the first stage, the information functions of items were calculated to place the items into panels as easy, medium, and difficult. The information functions for polytomous items were calculated using the GPCM, and for dichotomous items, they were calculated using the 2PLM. In the MST simulation, there are a total of six-panel designs: 1-2, 1-3, 1-4, 1-2-2, 1-2-3, and 1-3-3, and specific paths. Some sample panel designs and paths used in the study are shown in Figure 3.

Figure 3. *Examples of panel designs used in MST simulation.*



The creation of panels and test assembly process were carried out using the IBM CPLEX optimization program. Individuals were randomly assigned to panels in the study. The bottom-up test assembly method was used in creating panels. First of all, in order to maximize the module information, codes were written in the CPLEX program such that the item information levels were easy (θ = -1), low medium (θ =-0.3), medium (θ = 0), high medium (θ = 0.3) and difficult (θ = 1). Items that provided the highest information for each module were assigned to the modules based on the characteristics and numbers of each module. The selection of items designated as easy, medium, and difficult was carried out considering the total test length and the ratio of polytomous items. For example, the number of items in modules for the 1-2 panel design is detailed in Table 4.

Table 4. *Number of items in modules.*

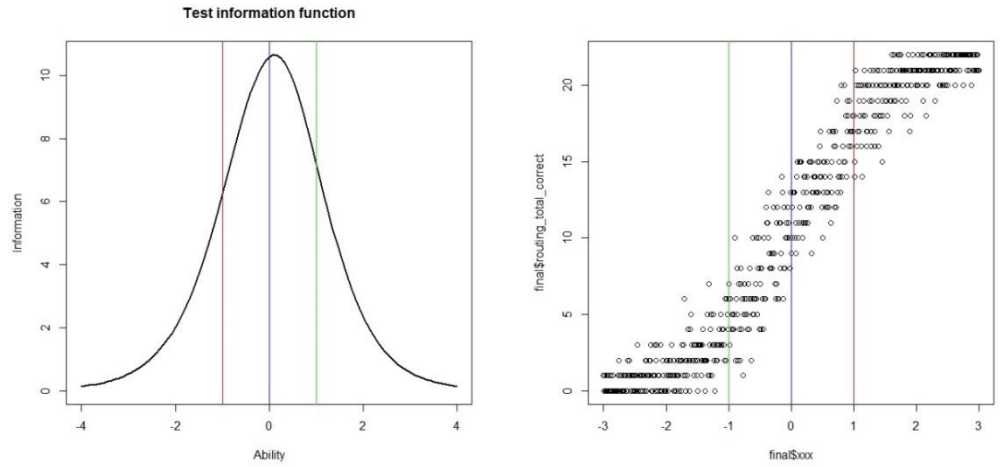
		Ratio of polytomous items								
		%10			%30			%50		
	Test length	20	40	60	20	40	60	20	40	60
Panel Design 1-2	1.Stage-M	1*/9	2*/18	3*/27	3*/7	6*/14	9*/21	5*/5	10*/10	15*/15
	2.Stage -E	1*/9	2*/18	3*/27	3*/7	6*/14	9*/21	5*/5	10*/10	15*/15
	2. Stage -H	1*/9	2*/18	3*/27	3*/7	6*/14	9*/21	5*/5	10*/10	15*/15

*Indicates the number of polytomous items in the modules. M: Medium, E: Easy, H: Hard

When examining Table 4, for the special condition of 1-2 design, 60 items and 10% polytomous item ratio, 30 items were selected and 10% (3 of them) were polytomous, giving high information at the point θ= 0. Similarly, under the condition θ=-1, 30 items were selected to provide high information and placed in the second easy module. Finally, under the condition θ=1, 30 items were selected to provide high information and placed in the second difficult module.

After creating modules and panels, MST simulations were continued with the R program. The simulation study for the MST was conducted using the 'mstR' package in R (Magis *et al*., 2018). The responses of individuals to all items were calculated using the GPCM. The EAP method was used for estimating abilities. In MST applications, commonly used routing methods include the NC (Weissman *et al*., 2007; Zenisky *et al*., 2010), ability estimation, and MFI (Weissman *et al*., 2007). NC routing method is an alternative method commonly used in IRT ability estimation in CAT (Armstrong, 2002). In this study, MFI and NC routing methods were used to transition from one stage to another. In the MFI routing method, the maximum amount of information is provided in one module and directed to the appropriate module in the next stage. In the NC-based routing method, an IRT model is used in the background to obtain a score distribution. Research indicates that the effectiveness of the NC directing method applied based on the raw number of correct responses is consistent with MST results designed for θ estimation (Armstrong, 2002). In the NC method, individuals are directed between stages based on the cumulative cut score point for each module. When determining the cut scores, the module information functions obtained by individuals according to MFI and their total number of correct responses were calculated for each stage. The correct numbers that provided the highest information were chosen as cut scores between modules. For example, for the 1-3 panel design with a 10% polytomous item ratio and a 40-item test, the graphs of the directing module information and total number of correct responses are shown in Figure 4.

Figure 4. *1-3 Panel design, 10% polytomous item ratio, routing module information and total numbers correct of the 40-item test.*



2.5.2. *Evaluation Criteria*

After obtaining simulation data under appropriate conditions for sub-problems, Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and correlations between actual and estimated ability levels were calculated to examine the accuracy of the obtained ability estimation. Bias is the square root of the average of the differences between the actual and estimated values. MAE is the absolute value of each difference between the actual and estimated values. RMSE is the square root of the average of the squares of the differences between the actual and estimated values. Low values of MAE and RMSE, and high correlation values indicate high accuracy of ability estimations. These values were calculated separately for each iteration for evaluation of the findings. Total values for each condition were divided by the number of iterations to report average values.

After obtaining MAE, RMSE, and correlation values, factorial ANOVA was conducted in SPSS using 100 iterations for each condition to determine if there was a significant difference among the condition variables. Effect sizes obtained from factorial ANOVA were examined. Three separate factorial ANOVA tests were conducted with MAE, RMSE, and correlation values as dependent variables, while all other conditions (ratio of polytomous items, test length, panel design, and routing method) were independent variables.

3. FINDINGS

3.1. Findings Related to Mean Absolute Error

The MAE findings for all conditions are shown in Table 5. The factorial ANOVA results on MAE are given in Table 6. Based on the overall findings, test length emerged as the most influential factor, accounting for 98% of the variance (η^2_p =.98), followed by the ratio of polytomous items in the test (η^2_p = .20). The routing method (MFI vs. NC) also showed a moderate effect (η^2_p = .13), while the panel design exhibited a smaller yet significant influence (η^2_p = .05). Differences across panel designs and routing methods were smaller but indicated a slight advantage for more complex panel designs and the MFI routing method. However, the ratio of polytomous items had minimal effect on MAE.

Table 5. *Findings of mean absolute error across all conditions.*

Routing Method	Panel Design	Ratio of Polytomous Items in the Test								
		%10			%30			%50		
		20	40	60	20	40	60	20	40	60
MFI	1-2	0.24	0.18	0.15	0.23	0.17	0.15	0.23	0.17	0.15
	1-3	0.24	0.18	0.15	0.23	0.17	0.15	0.23	0.17	0.15
	1-4	0.24	0.18	0.16	0.23	0.18	0.15	0.23	0.17	0.15
	1-2-2	0.23	0.18	0.15	0.23	0.17	0.14	0.23	0.17	0.14
	1-2-3	0.23	0.18	0.15	0.23	0.17	0.15	0.23	0.17	0.14
	1-3-3	0.23	0.18	0.15	0.23	0.17	0.15	0.23	0.17	0.14
NC	1-2	0.24	0.18	0.15	0.23	0.17	0.15	0.23	0.17	0.14
	1-3	0.24	0.18	0.16	0.24	0.18	0.15	0.24	0.18	0.15
	1-4	0.24	0.18	0.16	0.23	0.17	0.15	0.23	0.17	0.15
	1-2-2	0.25	0.18	0.16	0.24	0.17	0.15	0.24	0.17	0.16
	1-2-3	0.24	0.18	0.15	0.24	0.17	0.15	0.24	0.18	0.14
	1-3-3	0.24	0.18	0.16	0.24	0.18	0.15	0.25	0.18	0.15

Table 6. *ANOVA Results for grand mean absolute bias.*

Factor	Sum of Squares	*df*	Mean Square	*F*	*p*	η^2_p
Routing	0.05	1	0.05	1594.11	0.00	0.13
TL	14.00	2	7.00	228391.13	0.00	0.98
PIR	0.08	2	0.04	1297.90	0.00	0.20
PD	0.02	5	0.00	104.28	0.00	0.05
Routing*TL	0.01	2	0.00	112.19	0.00	0.02
Routing*PIR	0.00	2	0.00	18.52	0.00	0.00
Routing*PD	0.04	5	0.09	273.18	0.00	0.11
TL*PIR	0.00	4	0.00	10.24	0.00	0.00
TL*PD	0.01	10	0.00	24.61	0.00	0.02
PIR*PD	0.00	10	0.00	12.92	0.00	0.01
Routing*TL*PIR	0.00	4	0.00	12.23	0.00	0.01
Routing*TL*PD	0.01	10	0.00	32.32	0.00	0.03
Routing*PIR*PD	0.00	10	0.00	8.09	0.00	0.01
TL*PIR*PD	0.01	20	0.00	11.23	0.00	0.02
Routing*TL*PIR*PD	0.01	20	0.00	11.41	0.00	0.02
Error	0.33	10692	0.00			
Total	391.02	10800				

Note. PIR represents the Polytomous Item Ratio, TL refers to the Test Length, and PD denotes the Panel Design

3.2. Findings Related to Root Mean Square Error (RMSE)

The RMSE findings for all conditions are shown in Table 7. The factorial ANOVA results on this outcome are given in Table 8. The findings reveal that test length is the most influential factor (η^2_p=.98), with longer tests consistently reducing RMSE values across all conditions. For instance, RMSE values decreased notably as the test length increased from 20 to 60 items, irrespective of the routing method or panel design. The ratio of polytomous items, while statistically significant, exhibited a minimal impact on RMSE (η^2_p=.19). Routing methods demonstrated a moderate effect (η^2_p=.19), with MFI consistently outperforming NC routing in reducing RMSE, particularly in tests with longer lengths or more complex panel designs. Similarly, panel design contributed to slight improvements in estimation accuracy.

Table 7. *Findings of RMSE across all conditions.*

		Ratio of Polytomous Items in the Test								
		%10			%30			%50		
Routing Method	Panel Design	20	40	60	20	40	60	20	40	60
MFI	1-2	0.30	0.23	0.19	0.30	0.22	0.19	0.29	0.22	0.19
	1-3	0.30	0.23	0.19	0.30	0.22	0.19	0.29	0.22	0.19
	1-4	0.30	0.23	0.20	0.30	0.22	0.19	0.29	0.22	0.19
	1-2-2	0.30	0.22	0.19	0.29	0.22	0.18	0.29	0.22	0.18
	1-2-3	0.30	0.22	0.19	0.29	0.22	0.18	0.29	0.22	0.18
	1-3-3	0.30	0.22	0.19	0.29	0.22	0.18	0.29	0.22	0.18
NC	1-2	0.30	0.23	0.19	0.30	0.22	0.19	0.29	0.22	0.18
	1-3	0.30	0.23	0.20	0.30	0.23	0.19	0.30	0.22	0.19
	1-4	0.30	0.23	0.20	0.30	0.22	0.19	0.29	0.22	0.19
	1-2-2	0.32	0.24	0.20	0.31	0.22	0.20	0.31	0.22	0.20
	1-2-3	0.31	0.23	0.19	0.30	0.22	0.19	0.30	0.22	0.18
	1-3-3	0.31	0.23	0.21	0.31	0.22	0.19	0.32	0.23	0.19

Table 8. *ANOVA results for grand mean RMSE.*

Factor	Sum of Squares	*df*	Mean Square	*F*	*p*	η^2_p
Routing	0.11	1	0.11	2437.64	0.00	0.19
TL	22.72	2	11.36	251592.07	0.00	0.98
PIR	0.11	2	0.06	1219.26	0.00	0.19
PD	0.03	5	0.01	107.25	0.00	0.05
Routing*TL	0.02	2	0.01	191.05	0.00	0.04
Routing*PIR	0.00	2	0.00	17.60	0.00	0.00
Routing*PD	0.09	5	0.02	409.14	0.00	0.16
TL*PIR	0.00	4	0.00	6.33	0.00	0.00
TL*PD	0.02	10	0.00	38.50	0.00	0.04
PIR*PD	0.01	10	0.00	15.59	0.00	0.01
Routing*TL*PIR	0.00	4	0.00	13.39	0.00	0.01
Routing*TL*PD	0.02	10	0.00	50.33	0.00	0.05
Routing*PIR*PD	0.01	10	0.00	11.43	0.00	0.01
TL*PIR*PD	0.02	20	0.00	21.16	0.00	0.04
Routing*TL*PIR*PD	0.02	20	0.00	18.73	0.00	0.03
Error	0.48	10692	0.00			
Total	633.56	10800				

Note. See notes in Table 6.

3.3. Findings Related to Correlation

The correlation findings for all conditions are shown in Table 9. The factorial ANOVA results on this outcome are given in Table 10. As shown in Table 9, correlations increase with test length, consistently improving from 20 to 60 items, regardless of the routing method, panel design, or ratio of polytomous items. The ratio of polytomous items demonstrated minimal impact on correlation values, as observed in other metrics. Across all panel designs and routing methods, correlations remained virtually identical for tests with 10%, 30%, and 50% polytomous items. Routing methods showed slight differences, with MFI generally producing marginally higher correlations compared to NC routing, especially in longer tests. ANOVA reinforced that test length (η^2_p=.91) is the most influential factor in maximizing the alignment between estimated and true abilities, followed by the ratio of polytomous items (η^2_p=.07) and routing method (η^2_p=.07).

Table 9. *Findings of correlations across all conditions.*

		Ratio of Polytomous Items in the Test								
		%10			%30			%50		
Routing Method	Panel Design	20	40	60	20	40	60	20	40	60
MFI	1-2	0.95	0.97	0.98	0.96	0.98	0.98	0.96	0.98	0.98
	1-3	0.95	0.97	0.98	0.95	0.98	0.98	0.96	0.98	0.98
	1-4	0.95	0.97	0.98	0.95	0.97	0.98	0.96	0.98	0.98
	1-2-2	0.96	0.97	0.98	0.96	0.98	0.98	0.96	0.98	0.98
	1-2-3	0.96	0.98	0.98	0.96	0.98	0.98	0.96	0.98	0.98
	1-3-3	0.96	0.97	0.98	0.96	0.98	0.98	0.96	0.98	0.98
NC	1-2	0.95	0.97	0.98	0.96	0.98	0.98	0.96	0.98	0.98
	1-3	0.95	0.97	0.98	0.95	0.97	0.98	0.95	0.97	0.98
	1-4	0.95	0.97	0.98	0.96	0.98	0.98	0.96	0.98	0.98
	1-2-2	0.95	0.97	0.98	0.95	0.98	0.98	0.95	0.98	0.98
	1-2-3	0.95	0.97	0.98	0.95	0.98	0.98	0.95	0.97	0.98
	1-3-3	0.95	0.97	0.98	0.95	0.97	0.98	0.95	0.97	0.98

Table 10. *ANOVA results for correlation.*

Factor	Sum of Squares	*df*	Mean Square	*F*	*p*	η^2_p
Routing	0.01	1	0.01	810.48	0.00	0.07
TL	1.34	2	0.67	51715.76	0.00	0.91
PIR	0.01	2	0.01	389.50	0.00	0.07
PD	0.00	5	0.00	24.99	0.00	0.01
Routing*TL	0.01	2	0.00	244.91	0.00	0.04
Routing*PIR	0.00	2	0.00	14.16	0.00	0.00
Routing*PD	0.01	5	0.00	106.40	0.00	0.05
TL*PIR	0.01	4	0.00	99.69	0.00	0.04
TL*PD	0.00	10	0.00	24.70	0.00	0.02
PIR*PD	0.00	10	0.00	7.57	0.00	0.01
Routing*TL*PIR	0.00	4	0.00	3.54	0.01	0.00
Routing*TL*PD	0.00	10	0.00	34.33	0.00	0.03
Routing*PIR*PD	0.00	10	0.00	7.16	0.00	0.01
TL*PIR*PD	0.00	20	0.00	4.63	0.00	0.01
Routing*TL*PIR*PD	0.00	20	0.00	3.21	0.00	0.01
Error	0.14	10692	0.00			
Total	10158.54	10800				

Note. See notes in Table 6.

4. DISCUSSION and CONCLUSION

This study investigates the impact of varying proportions of polytomous items, test length, panel design, and routing methods on the accuracy of ability estimation in MST. The findings highlight the critical role of test length and the ratio of polytomous items in improving estimation accuracy, with secondary influences from panel design and routing methods. These results contribute to a nuanced understanding of the conditions under which MST achieves optimal precision.

The results demonstrate that as the ratio of polytomous items in the test increases from 10% to 50%, the estimation of abilities improves, reflected by lower MAE and RMSE values and higher correlations between estimated and true abilities. This aligns with the theoretical premise that polytomous items provide more information per item than dichotomous ones (Embretson & Reise, 2000). The enhanced test information associated with a higher ratio of polytomous items enables more precise ability estimation, a finding corroborated by previous research on mixed-format tests (Kim *et al*., 2012). However, the differences between 10% and 30% proportions were relatively minor, suggesting that the benefits of polytomous items become more pronounced at higher ratios. In contrast to Park (2015), who found that an increased ratio of polytomous items reduced measurement precision due to challenges in test construction, this study observed consistent improvements in accuracy, likely attributable to well-constructed test panels that maintained sufficient test information.

Test length emerged as the most significant factor influencing ability estimation, as indicated by the ANOVA results, which showed that it explained the highest variance across MAE, RMSE, and correlations. Longer tests consistently yielded lower error rates and higher correlations, with tests comprising 40 or 60 items achieving the most accurate ability estimates. This finding is consistent with prior MST studies, which emphasize the role of test length in enhancing measurement precision (Erdem-Kara, 2019; Kim *et al*., 2013; Patsula, 1999). The results underscore the importance of prioritizing test length during test assembly to achieve optimal accuracy.

Panel design also influenced estimation accuracy, with three-stage panel designs generally outperforming two-stage designs. This result aligns with the broader MST literature, which indicates that additional stages in the panel design allow for more refined routing decisions, thereby reducing estimation error (Dogruöz, 2018; Patsula, 1999). However, the effect size was smaller compared to the test length and the ratio of polytomous items, suggesting that panel design plays a supportive but less critical role in overall accuracy.

Routing methods showed a modest but significant effect on estimation accuracy, with the MFI method consistently outperforming the NC method. This difference was particularly evident in conditions involving longer tests and higher proportions of polytomous items. Unlike Weissman *et al*. (2007), who found comparable performance between MFI and NC routing methods in terms of classification accuracy, this study suggests that MFI provides superior measurement accuracy in contexts requiring precise ability estimation. The superior performance of MFI in this study may stem from its ability to optimize information across items, whereas NC routing is more constrained by cutoff score determination.

Overall, this study highlights the critical interplay among test length, item composition, panel design, and routing methods in achieving accurate ability estimates in MST. The results emphasize the importance of prioritizing test length and leveraging the informational advantages of polytomous items while optimizing panel design and routing methods to further enhance precision. These findings provide practical guidance for test developers in designing MST frameworks that balance psychometric precision with operational constraints. Future research should explore these dynamics further in operational settings and extend the analysis to other adaptive testing contexts.

4.1. Recommendations

The findings of this study provide several practical and theoretical insights for optimizing MST frameworks. Key recommendations are presented in the following paragraphs:

First, test length emerged as the most influential factor in improving ability estimation accuracy. Therefore, it is recommended to prioritize longer tests in MST applications, particularly when measurement precision is critical. Tests with 40 or more items consistently yielded lower error rates and higher correlations, regardless of other factors. However, operational constraints such as time limitations and test-taker fatigue should also be considered when extending test lengths.

Second, the ratio of polytomous items in mixed-format tests significantly enhances the accuracy of ability estimates. While the effects of increasing polytomous item ratios from 10% to 30% were modest, the benefits became more pronounced at a 50% ratio. These findings suggest that incorporating a balanced ratio of polytomous items can maximize test information and improve measurement precision. However, test developers should ensure that the item pool remains sufficiently diverse and aligned with test specifications to maintain content validity.

Third, the panel design should be tailored to the specific conditions of the test. Under conditions with long tests (60 items) and high proportions of polytomous items (50%), differences in measurement accuracy across panel designs were minimal. This suggests that simpler designs (e.g., "1-2") could be employed in such cases to reduce operational complexity without compromising accuracy. For tests with shorter lengths or lower proportions of polytomous items, more complex designs (e.g., "1-2-3") may be beneficial to enhance measurement precision.

Fourth, the choice of routing method is critical for ensuring accuracy in ability estimation. The MFI routing method consistently outperformed the NC method across most conditions, particularly for longer tests and higher proportions of polytomous items. However, interactions between panel designs and routing methods were observed, such as with the "1-2" and "1-4" designs, where both methods performed similarly. In such cases, operational considerations,

such as ease of implementation or alignment with institutional goals, may guide the choice of routing method.

Future research should address the limitations of this study by expanding the scope of item parameters and test conditions. For instance, the data in this study were based on item parameters from the PISA 2018 Reading Skills domain test. Investigating other domains or disciplines with different item characteristics (e.g., difficulty, discrimination) could provide broader insights. Additionally, exploring how MST frameworks perform across diverse cultural and linguistic contexts would offer valuable perspectives on the generalizability of these findings.

Finally, this study assumed that individuals' ability levels followed a normal distribution. Future research could investigate how alternative distributions (e.g., uniform, skewed) influence ability estimation accuracy. Additionally, examining the effects of systematically arranging polytomous items across modules, rather than random distribution, could yield insights into how item pool characteristics interact with panel designs and routing methods. Such investigations could lead to more tailored and flexible MST designs that accommodate diverse testing needs. These recommendations provide actionable guidance for test developers while identifying key areas for future research to enhance the flexibility, accuracy, and fairness of MST applications.

Acknowledgments

This study was produced from the doctoral thesis of the first 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

Hasibe Yahsi Sari: Literature Review, Investigation, Data Collection, Visualization, Statistical Analysis, and Writing. **Hulya Kelecioglu**: Supervision, Methodology, and Writing.

Orcid

Hasibe Yahsi Sari https://orcid.org/0000-0002-0451-6034
Hulya Kelecioglu https://orcid.org/0000-0002-0741-9934

REFERENCES

Armstrong, R.D. (2002). *Routing rules for multi-form structures* (LSAC Computerized Testing Report No. 02-08). Law School Admission Council. https://searchworks.stanford.edu/view/6794803

Bock, R.D., & Mislevy, R.J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, *6*(4), 431-444. https://doi.org/10.1177/014662168200600405

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

Cıkrıkcı, N., Yalçın, S., Kalender, İ., Gül, E., *et al*. (2020). Development of a computerized adaptive version of the Turkish Driving Licence Exam. *International Journal of Assessment Tools in Education*, *7*(4), 570-587. https://doi.org/10.21449/ijate.716177

Dallas, A. (2014). *The effects of routing and scoring within a computer adaptive multi-stage framework* [Doctoral dissertation, University of North Carolina]. University of North Carolina Libraries. https://libres.uncg.edu/ir/uncg/f/Dallas_uncg_0154D_11394.pdf

Doğruöz, E. (2018). *Bireyselleştirilmiş çok aşamalı testlerin test birleştirme yöntemlerine göre incelenmesi* [Doktora tezi, Hacettepe Üniversitesi]. Hacettepe Üniversitesi Açık Erişim Sistemi. http://hdl.handle.net/11655/5298

Erdem-Kara, B. (2019). *Değişen madde fonksiyonu gösteren madde oranının bireyselleştirilmiş bilgisayarlı ve çok aşamalı testler üzerindeki etkisi* [Doktora tezi, Hacettepe Üniversitesi]. Hacettepe Üniversitesi Açık Erişim Sistemi. http://hdl.handle.net/11655/11968

Han, K.T. (2007). WinGen: Windows Software That Generates Item Response Theory Parameters and Item Responses. *Applied Psychological Measurement*, *31*(5), 457-459. https://doi.org/10.1177/0146621607299271

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

Kim, J., Chung, H., & Dodd, B.G. (2010, April). *Comparing routing methods in the multistage test based on the partial credit model* [Conference presentation] Annual meeting of the American Educational Research Association, Denver, CO.

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/10.1177/001316441142897

Kim, J., Chung, H., Park, R., & Dodd, B.G. (2013). A comparison of panel designs with routing methods in the multistage test with the partial credit model. *Behavior Research Methods, 45*(4), 1087-1098. https://doi.org/10.3758/s13428-013-0316-3

Kim, J., & Dodd, B.G. (2014). Mixed-format multistage tests: Issues and methods. In D. Yan, A.A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing: Theory and applications* (pp. 55–67). CRC Press.

Kim, H., & Plake, B.S. (1993). Monte carlo simulation comparison of two-stage testing and computerized adaptive testing. *Annual Meeting of the National Council on Measurement in Education*, Atlanta, GA. https://eric.ed.gov/?id=ED357041

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

Luecht, R.M., & Nungester, R.J. (1998). Some practical examples of computer-adaptive sequential testing. *Journal of Educational Measurement*, *35*(3), 229-249. http://www.jstor.org/stable/1435202

Magis, D., Yan, D., & von Davier A.A. (2018). *Package 'mstR': Procedures to generate patterns under multistage testing*. https://cran.r-project.org/web/packages/mstR/mstR.pdf

OECD (2002). *Proposed standard practice for surveys on research and experimental development: Frascati Manual 2002*. OECD Publishing. https://doi.org/10.1787/9789264199040-en

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]. University of Texas Libraries. http://hdl.handle.net/2152/31011

Park, R., Kim, J., Chung, H., & Dodd, B.G. (2014). Enhancing pool utilization in constructing the multistage test using mixed-format tests. *Applied Psychological Measurement*, *38*(4), 268-280. https://doi.org/10.1177/0146621613515

Park, R., Kim, J., Chung, H., & Dodd, B. G. (2017). The development of mst test information for the prediction of test performances. *Educational and Psychological Measurement*, *77*(4), 570–586. https://doi.org/10.1177/0013164416662960

Patsula, L.N. (1999). *A comparison of computerized adaptive testing and multistage testing.* [Doctoral dissertation, University of Massachusetts]. University of Massachusetts Libraries. https://doi.org/10.7275/10994910

R Core Team. (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. http://www.R-project.org/

Rosa, K., Swygert, K.A., Nelson, L., & Thissen, D. (2001). Item response theory applied to combinations of multiple-choice and constructed-response items—scale scores for patterns of summed scores. In D. Thissen & H. Wainer (Eds.), *Test scoring* (1st ed., pp. 253–292). Routledge. https://doi.org/10.4324/9781410604729

Sahin, M.G., & Ozturk, N.B. (2019). Analyzing the maximum likelihood score estimation method with fences in ca-MST. *International Journal of Assessment Tools in Education*, *6*(4), 555-567. https://dx.doi.org/10.21449/ijate.634091

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

Senel, S., & Kutlu, O. (2018). Computerized adaptive testing design for students with visual impairment. *Education and Science Journal*,*43*(194), 261-284. https://doi.org/10.15390/EB.2018.7515

Wang, K. (2017). *A fair comparison of the performance of computerized adaptive testing and multistage adaptive testing* [Doctoral dissertation, University of Michigan State]. University of Michigan State Libraries. https://doi.org/doi:10.25335/ypy5-6g68

Weiss, D.J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement*, *6*(4), 473-492. https://doi.org/10.1177/014662168200600408

Weiss, D.J. (1983). Latent trait theory and adaptive testing. In Weiss D.J. (Ed.), *New horizons in testing* (pp. 5-7). Academic Press.

Weiss, D.J., & Kingsbury, G.G. (1984). Application of computer adaptive testing to educational problems. *Journal of Educational Measurement*, *21*(4), 361-375. https://doi.org/10.1111/j.1745-3984.1984.tb01040.x

Weissman, A., Belov, D.I., & Armstrong, R.D. (2007). Information-based versus number-correct routing in multistage classification tests, *LSAC Computerized Testing Report, 7(*5). Law School Admission Council.

Yahsi Sari, H., & Kelecioglu, H. (2023). Ability estimation with polytomous items in computerized multistage tests. *Journal of Measurement and Evaluation in Education and Psychology*, *14*(3), 171-184. https://doi.org/10.21031/epod.1056079

Yamamoto, K., Shin, H., & Khorramdel, L. (2019). Introduction of multistage adaptive testing design in PISA 2018. *OECD Education Working Papers, No. 209*, OECD Publishing. https://doi.org/10.1787/b9435d4b-en

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]. University of Massachusetts Libraries. https://doi.org/10.7275/18739572

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

Zheng, Y., Nozawa, Y., Gao, X., & Chang, H.H. (2012). Multistage adaptive testing for a large-scale classification test: The designs, automated heuristic assembly, and comparison with other testing modes. *ACT Research Reports 2012-6.* https://www.act.org/content/dam/act/unsecured/documents/ACT_RR2012-6.pdf