ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2023; 14(3); 171-184

Ability Estimation with Polytomous Items in Computerized Multistage Tests

Hasibe YAHŞİ SARI* Hülya KELECİOĞLU**

Abstract

This study aims to examine how individuals' ability estimations change under different conditions in tests consisting of polytomous items in a computerized multistage test environment. In this simulation study, 108 (3x3x6x2=108) conditions were examined, consisting of three categories (3, 4, and 5), three test lengths (10, 20, and 30), six-panel designs (1-2, 1-2-2, 1-3, 1-3-3, 1-4, and 1-4-4), and, two routing methods (Maximum Fisher Information (MFI) and Random). Simulations and analyses were carried out in the mstR package in the R program, with a pool of 200 items, 1000 people, and 100 replications (i.e. iterations). The mean absolute bias, RMSE, and correlation values were calculated as the research outcomes. This study discovered that as the number of categories and test lengths increase, the mean absolute bias and RMSE values decrease, while the correlation values increase. Although MFI and random methods have similar tendencies regarding routing methods, MFI provides better results. Furthermore, there is a similarity between the panel designs in terms of results.

Keywords: Computerized multistage tests, polytomous items, routing method.

Introduction

Traditional paper-and-pencil tests have been replaced by computerized adaptive tests (CAT) in educational and psychological institutions. CATs are the tests in which the abilities of individuals are estimated with a scaled item pool before the exam, which has rules of starting, progressing, and ending according to the individual's previously known or predicted ability (Weiss, 1982). There are many advantages to CATs compared to traditional paper-and-pencil applications. For instance, an advantage of CATs is the increased accurate ability estimation by using fewer items and prompt disclosure of results (Weiss, 1983). However, CAT applications also have disadvantages such as different test lengths (i.e., fixed-length is also available), different questions being asked, and the individual not being able to return to the previous question. Due to the overwhelming disadvantages of CAT, the use of computerized multistage tests (MST) is becoming widespread (Hendrickson, 2007; MacGregor et al., 2022; Zenisky et al., 2009).

MST combines the advantages of CAT and paper-pencil tests. MST achieves this by adjusting the tests based on each individual. While CATs are adapted to the individual at the item level, MSTs are adapted to the individual at the module level (Zenisky et al., 2009). Unlike CATs, MSTs consist of item groups called modules and stages. Modules consist of items; stages consist of modules; panels consist of stages. MSTs provide the opportunity to move between the items in the module and allow test preparers to better control the test content compared to CATs (Hendrickson, 2007; Sari et al., 2016).

The characteristics of the item pool are important in MSTs, as in CATs. Unlike CATs, MSTs have their own terminology including panel structure, routing method, module, and stage. A module consists of a group of items at the same or similar difficulty level. A stage consists of a different number of modules at different difficulty levels such as easy, medium or hard modules. A panel design is comprised of

* Teacher, Ministry of National Education, Kilis-Türkiye, hsbyahsi@gmail.com, ORCID ID: 0000-0002-0451-6034
** Prof.Dr., Hacettepe University, Faculty of Education, Ankara-Türkiye, hulyakelecioglu@gmail.com, ORCID ID: 0000-0002-0741-9934

To cite this article:
Yahşi Sarı, H. & Kelecioğlu ,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

Received: 10.01.2022
Accepted: 21.12.2022

different stages. Test assembly is the process of building modules, stages, and panels so it is one of the most important steps in an MST.

An MST functions as follows: Individuals take the first-stage module, called the routing module. Then, the individual is selected for the appropriate module based on the current ability level at the second stage. Finally, the exam continues until a test taker completes all required stages.

Background and Literature Review

Various past studies on routing methods generally apply Approximate Maximum Information (AMI), Defined Population Intervals (DPI), and convergent and random routing methods (Kim et al., 2010; Zenisky, 2004). Routing methods that are based on IRT are other frequently used kinds. These methods are Maximum Fisher Module Information (MFI), Maximum Likelihood Weighted Module Information (MLWMI), Maximum Posterior Weighted Module Information (MPW MI), Maximum Module Kullback-Leibler Information (MKL), Maximum Posterior Module Kullback-Leibler Information (MKLP) and random. In this study, MFI and random routing methods were used. The MFI routing method is based on the item information level. In MST, routing with MFI is made to the next stage according to the cumulative information obtained from the module items. The MFI routing method directs individuals to the module, explaining their ability levels to the maximum (Weissman et al., 2007). In the random routing method, theta estimation is made after the module is taken in the routing module. Then, the individual is randomly assigned to a module in the next stage. On the other hand, individuals are referred to any of the following stage modules with equal probability, regardless of their scores in a previous stage.

One of the conditions of MST is panel design. A panel design is formed by the combination of different numbers of modules and stages. Panel design may vary depending on the purpose of the MSTs. For example, 1-3 panel patterns consist of 2 stages and four modules. There is 1 module in the first stage (also called the routing module) and three in the 2nd stage. In a 1-3 panel design, the difficulty levels of the modules are usually determined as easy, medium, and complicated in the 2nd stage. 1-2, 1-2-2, 1-3, 1-3-3, 1-4, and 1-4-4 panel designs, which are preferred in the literature, were used in this study (Kim et al., 2010; Oztürk, 2019; Sarı & Raborn, 2018).

It is known that test length affects ability estimation in MST designs (Luecht, 2000; Sarı & Raborn, 2018). Based on the literature, while some studies use different numbers of items at all stages (Macken-Ruiz, 2008), some other studies use the same number of items (Kim et al., 2013). Kim et al. (2013) compared the MST designs that they created based on the partial credit model, using different routing methods and panel designs, in the context of the classification test. As a result, it was observed that the accuracy of the ability estimations increased as the test length increased. Previous studies using polytomous test items mainly used 9-20 items (Chen, 2010; Kim et al., 2013; Macken-Ruiz, 2008). Based on the studies examined in the literature, 10, 20, and 30 test lengths were examined in this study.

MST applications are made with dichotomous (i.e. binary) and polytomous items. Zenisky (2004) compared various panel designs with different routing methods (DPI, proximity, and random) to estimate the ability and determine its precision. The item pool was based on the three-parameter logistic IRT model. Several studies in the literature examine the ability estimations of MST designs using two-category (i.e. binary) data using different conditions and routing methods (Oztürk, 2019; Sarı & Raborn, 2018; Zenisky, 2004). Polytomous items provide more information, allowing more accurate findings in ability estimation (Donoghue, 1994). However, few research studies use different routing methods in polytomous data. Studies in the literature which use polytomous items are generally designed according to the partial credit model (Kim et al., 2010; Kim et al., 2013). Nonetheless, GPCM is used in current studies and applications such as PISA 2018 (Choi, & Asilkalkan, 2019; Ridho, 2022). Thus, in this study, we utilized GPCM when generating and analyzing polytomous items.

This study is unique because it was designed with different panel designs, routing methods, and items produced according to the generalized partial credit model. In addition, AMI, DPI, M-AMI, M-DPI, SL-DPI, and ML-DPI routing methods are frequently used in the literature (Kim et al., 2010; Kim et al.,

2013; Zenisky, 2004). Some studies use MFI, MLWMI, MPWMI, MKL, MKLP, and random routing methods with dichotomous items (Oztürk, 2019; Sari & Rabon, 2018). Also, in the new MSTGen data generator program developed by Han (2022), there are three options for the routing methods: MFI, matching b-value, and random. The MFI routing method selects the most informative item with the highest accuracy due to its formulation (Luo et al., 2016). Although MFI and random are essential methods that have been frequently used (Svetina et al., 2019), there is no study in which one performs better in polytomous items. The results of this study will provide essential contributions in terms of being a guide to the optimum conditions of real applications that are likely to be applied in the future.

In this study, we researched the answer to the following question presented: "In computerized adaptive multistage tests, in tests consisting of polytomous items (3, 4, and 5 categories), how do the ability estimations of individuals change depending on test length (10, 20, and 30), panel designs (1-2, 1-2- 2, 1-3, 1-3-3, 1-4, and 1-4-4) and routing methods (Maximum Fisher Information [MFI] and Random)?"

Methods

This research is a simulation study, and the aim of the study is to examine the effects of simulation conditions (e.g., test length, number of item categories, panel design, and routing method) on ability estimation under the context of having polytomous items. Within the scope of the research, three categories (3, 4, and 5), three test lengths (10, 20, and 30), six-panel designs (1-2, 1-2-2, 1-3, 1-3-3, 1-4, and 1-4-4) and two routing methods (Maximum Fisher Information [MFI] and Random), 108 (3x3x6x2) conditions were examined. The conditions of the study are shown in Table 1.

Table 1

Simulation Conditions

Condition	Number of Levels	Levels
Number of Category	3	3-category 4-category 5-category
Test Length	3	10 items 20 items 30 items
Panel Design	6	1-2 1-2-2 1-3 1-3-3 1-4 1-4-4
Routing Method	2	MFI Random
Total	3x3x6x2=108	

Sample size (1000), sample ability distribution [N(0,1)], item pool size (200 items), and ability estimation method (Expected a priori-EAP) were kept constant in the study. 100 iterations were run for each condition.

Three separate item pools, each consisting of 200 items in 3, 4, and 5 categories to be used in the research, were generated with the WinGen (Han, 2007) program. Item parameters were produced according to 208 items' descriptive statistics consisting of 3, 4, and 5 categories as Macken Ruiz (2008) used in his dissertation. When generating *a* and *b* parameters under different numbers of item categories (e.g., 3, 4, and 5-category), we used a uniform distribution. The parameter *a* was in the range of [0.68, 1.5] for 3-category items, [0.57, 1.01] for 4-category items, and [0.54, 1] for 5-category items. The *b* parameter was between [-2.77, 3.41] for 3-category items, [-3.01, 3.44] for 4-category items, and [-3.15,

1.68] for 5-category items. With the simulation, 200 polytomous items were produced according to the generalized partial score model (GPCM) (Muraki, 1992). The GPCM formulation is as follows (Embretson & Reise, 2013):

$$P_{ix} = \frac{\exp[\sum_{j=0}^{x} a_i\,(\theta - g_{ij})]}{\sum_{r=0}^{m} exp\,[\sum_{j=0}^{r} a_i\,(\theta - g_{ij})]} \quad (1)$$

where m is the number of categories, x is the student's score on the item, i is the item index, θ is the student's ability, a is discrimination parameter for the item j. Substituting the category information function a simplified equation for polytomous item information is calculated as (Samejima, 1969; Dodd et al., 1995):

$$I_i(\theta_j) = \sum_{x=0}^{m_i} \frac{[P'_{ix}(\theta_j)]^2}{P_{ix}(\theta_j)} \quad (2)$$

Descriptive statistics of item parameters are given in Table 2.

Table 2

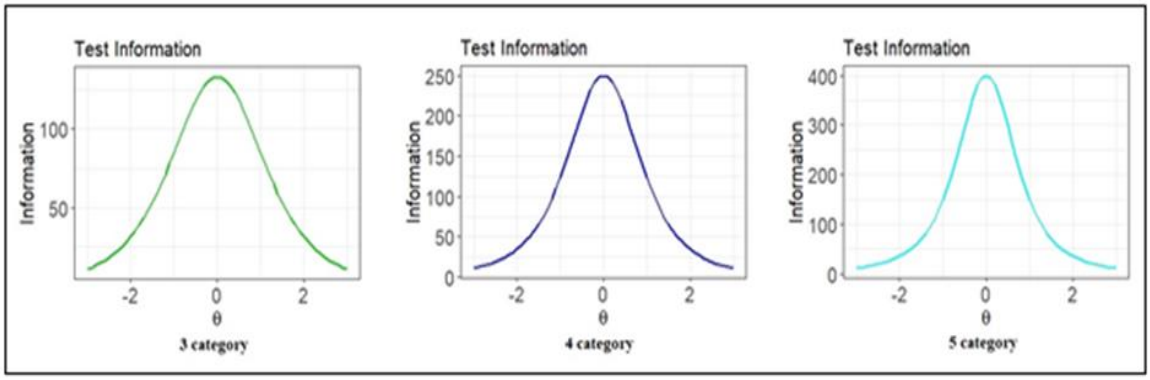
Descriptive Statistics For The Item Parameters Across The Condition

	3-Category			4-Category				5-Category				
Statistics	*a*	b_1	b_2	*a*	b_1	b_2	b_3	*a*	b_1	b_2	b_3	b_4
Min.	0.68	-2.77	-1.63	0.57	-3.01	-1.73	-0.86	0.54	-3.15	-2.31	-1.47	-0.87
Max.	1.50	0.94	3.41	1.01	-0.81	2.35	3.44	1.00	-1.05	1.45	1.68	1.03
Mean	1.09	-0.63	0.49	0.78	0.75	-0.01	0.85	0.78	0.94	-0.28	0.33	3.03

Item information functions were calculated in R program (R Development Core Team, 2018), and modules and panels were built in IBM CPLEX program (ILOG, 2006). Cplex is a mathematical modeling program that solves optimization problems consisting of linear or quadratic equations with the most precise results possible. The Cplex program selected the most appropriate items to be placed in each module from the item pool. Figure 1 shows test information function graphs of three different item pools consisting of 3, 4, and 5-category items.

Figure 1

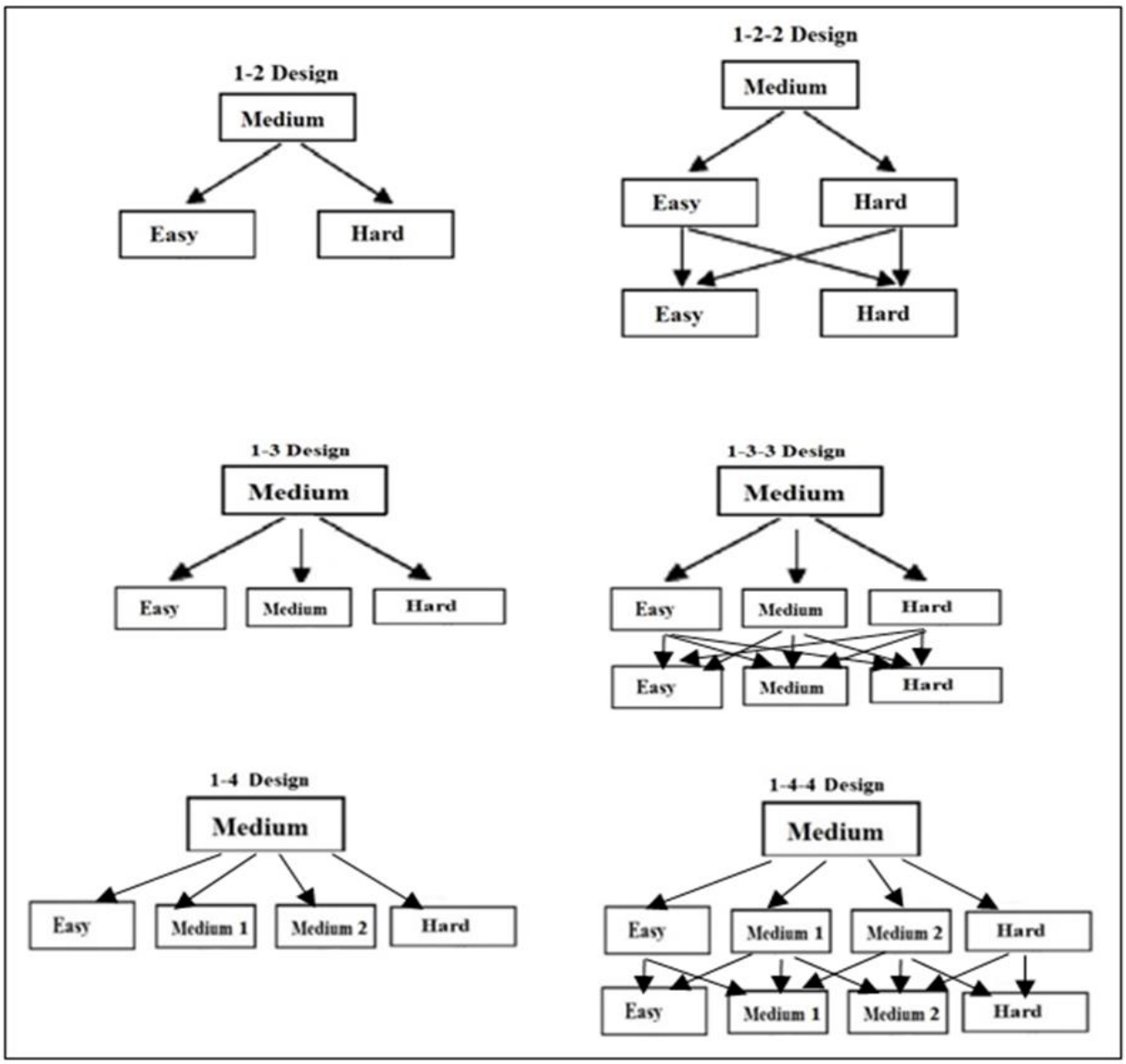
Test Information Functions of Item Pools



The routing module comprises items with medium difficulty levels. For items of medium difficulty, the total item information is maximized at theta level of 0. In 1-2, 1-3, 1-2-2, and 1-3-3 panel designs, easy modules are composed of items with easy difficulty levels meaning that module-level total item information is maximized at the theta level of -1. Lastly, hard modules are composed of items with hard difficulty levels meaning that module-level total item information is maximized at the theta level of +1. In 1-4 and 1-4-4 panel designs, the routing module comprises items with medium difficulty levels, as in the other panel designs. Easy modules are composed of items with easy difficulty levels. Lastly, hard modules are composed of items with hard difficulty levels. As the panel design implies, in 1-4 and 1-4-4 panel designs, there are four modules at different difficulty levels at other stages. These modules are easy, medium-1, medium-2, and hard. For the easy modules, module-level total item information is maximized at the theta level of -1. For the medium -1 module, module-level total item information is maximized at the theta level of -0.33. For the medium-2 modules, module-level total item information is maximized at the theta level of +0.33. For the hard modules, module-level total item information is maximized at the theta level of +1. All panel designs used in the study are shown in Figure 2.

Figure 2

All Panel Designs Used In The Study



As we mentioned above, the sample size is 1000, and there are 108 conditions in this study. The mean absolute bias, RMSE, and correlation values were obtained with a total of 10.800 iterations, 100 iterations for each condition. Four-way ANOVA was run in SPSS for the results. F values and partial η^2 statistics were used to determine the significance of the effects of the factors. Obtained results are given in the findings section.

The research conditions were determined by examining the literature, and taking into account the most frequently used conditions in simulations and real applications (see Rutkowski et al., 2022; Svetina et al., 2019). The studies in the literature related to the conditions in this study are explained in detail in the literature review section. Mean absolute bias (MAB), mean squares of error (RMSE), and correlation values were calculated to evaluate the results. These statistics were calculated from the following formulas.

The bias is the average of the difference between the actual and the predicted value. The bias ($\bar{e}$) is formulated as follows:

$$\bar{e} = \sqrt{\frac{\sum_{J=1}^{N}(\widehat{\theta_J} - \theta_J)}{N}} \tag{3}$$

The mean absolute error (MAE) derives from the unaltered magnitude (absolute value) of each difference.

$$Mean\ Absolute\ Bias = [n^{-1} \sum_{i=1}^{n} |\widehat{\theta_J} - \theta_J|] \tag{4}$$

The RMSE is the mean of the squared difference between the actual and predicted value. The mean squared error is formulated as follows.

$$RMSE = \sqrt{\frac{\sum_{J=1}^{N}(\widehat{\theta_J} - \theta_J)^2}{N}} \tag{5}$$

The correlation between actual (θ) and calculated ($\hat{\theta}$) skill levels ($p(\widehat{\theta_J}, \theta_J)$) is formulated as follows.

$$p(\widehat{\theta_J}, \theta_J) = \frac{cov(\widehat{\theta_J}, \theta_J)}{\sigma_{\widehat{\theta_J}}\, \sigma_{\theta_J}} \tag{6}$$

Results

Overall, when we analyzed the findings in terms of panel design, panel designs 1-2, 1-3, 1-4, 1-2-2, 1-3-3, and 1-4-4 produced very similar results under different conditions. However, the routing method and several item categories changed the study outcomes. Therefore, the study findings regarding routing methods and the number of categories were discussed.

Mean Absolute Bias

Table 3 shows the mean absolute bias values obtained under all simulation conditions. Regardless of panel design, number of categories, and test length, MFI gives better results than random routing

methods. Figure 3 shows the graphs of the mean absolute bias values according to the number of categories. The mean absolute bias decreased as the test length increased. Under the same conditions, as the number of categories changed from 3 to 4, there was a slight increase in the mean absolute bias values. However, MST conditions consisting of 5-category items had the lowest mean absolute bias values. The lowest mean absolute bias is seen in the MFI routing method (.149) in the 5-category, 30-item test, and 1-3-3 panel design. The highest mean absolute bias is seen in the random routing method in the tests in 4-category, 10-item, and 1-2-2 panel designs (.301). The highest score is highlighted in bold, and the lowest score is marked in bold and italic in Table 3.

Table 3

Findings of Average Absolute Bias Across All Conditions

Routing Method	Panel Design	3-Category			4-Category			5-Category		
		10	20	30	10	20	30	10	20	30
MFI	1-2	.261	.195	.164	.277	.207	.174	.242	.179	.150
	1-2-2	.259	.194	.163	.276	.205	.173	.241	.178	.150
	1-3	.260	.194	.165	.278	.207	.176	.240	.177	.150
	1-3-3	.257	.192	.164	.275	.206	.174	.239	.176	***.149***
	1-4	.259	.197	.165	.277	.208	.177	.241	.178	.153
	1-4-4	.256	.192	.165	.278	.207	.178	.237	.178	.151
Random	1-2	.295	.223	.186	.300	.225	.188	.261	.195	.162
	1-2-2	.295	.224	.186	**.301**	.226	.189	.261	.196	.163
	1-3	.292	.221	.186	.299	.224	.190	.260	.194	.162
	1-3-3	.294	.222	.188	.299	.226	.192	.259	.197	.164
	1-4	.293	.225	.189	.299	.225	.191	.262	.197	.165
	1-4-4	.296	.226	.193	.300	.230	.194	.264	.198	.165

Figure 3

Average Absolute Bias Values According to The Number of Categories

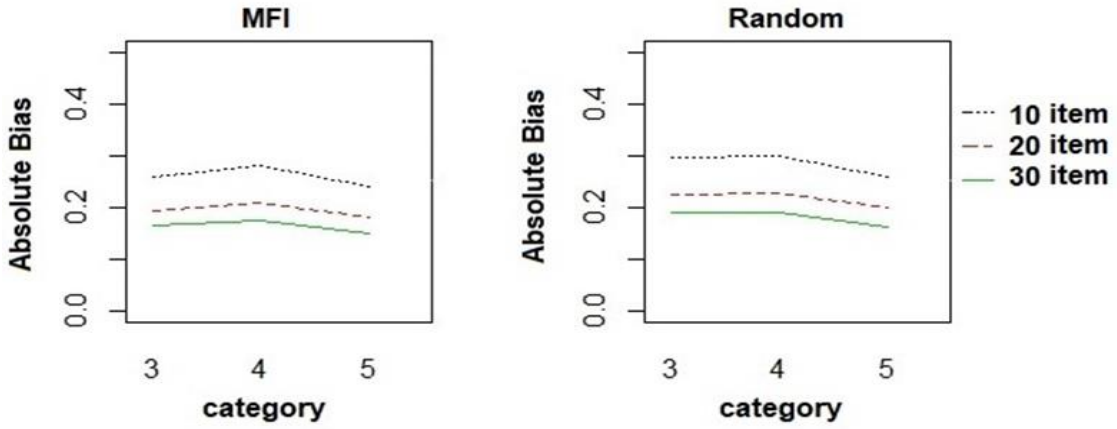


Table 4 shows that ANOVA results for mean absolute bias indicate that most interaction and main effects were significant. Four factors ANOVA was significant (η^2 =.922). However, the factors with the highest partial η^2 were the main effects of routing and test length (η^2 =.927). These effects explained about %93 of the variance in the mean absolute bias. The main effects of test length (η^2=.868) was the factor with the next largest partial η^2. The factor explained about 87% of the variance in the mean absolute bias. When category and panel design were added to routing and test length separately, the factor explained about 83% of the variance in the mean absolute bias (η^2 =.827).

Tablo 4

ANOVA Results for Grand Mean Absolute Bias

Factor	Sum of Squares	df	Mean Square	F	p	η^2_p
Routing	1.401	1	1.401	43657.236	.000	.803
Test Lenght	2.250	2	1.125	35054.378	.000	.868
Category	1.376	2	.688	21435.204	.000	.800
Panel Design	.344	5	.069	2142.806	.000	.501
Routing * Test Lenght	4.328	2	2.164	67414.127	.000	.927
Routing * Category	.937	2	.469	14599.161	.000	.732
Routing * Panel Design	.120	5	.024	746.228	.000	.259
Test Lenght* Category	.252	4	.063	1962.405	.000	.423
Test Lenght* Panel Design	.861	10	.086	2683.301	.000	.715
Category * Panel Design	.593	10	.059	1848.531	.000	.634
Routing*Test Lenght * Category	1.644	4	.411	12806.986	.000	.827
Routing * Test Lenght * Panel design	1.644	10	.164	5122.090	.000	.827
Routing * Category *Panel Design	.968	10	.097	3015.402	.000	.738
Test Lenght* Category * Panel design	1.200	20	.060	1869.315	.000	.778
Routing * Test Lenght * Category * Panel Design	4.083	20	.204	6360.937	.000	.922
Residuals	.343	10692	.000			
Total	528.233	10800				

Root Mean Square Error

Table 5 shows the RMSE values obtained under all research conditions. Figure 4 shows the graphs of RMSE values according to the number of categories. Regardless of panel pattern, number of categories, and test length, MFI gives better RMSE results than the random routing method. As the test length increased, the RMSE value decreased in both routing methods. As the number of categories increased, the RMSE value decreased. The lowest RMSE value is seen in the 5-category, 30-item test, in 1-3-3 panel design, in the MFI routing method (.190). The highest RMSE values are seen in the random routing method (.383) in the 10-item test with 3 and 4 categories. The highest RMSE value is for 4 categories in 1-2-2 panel design. Another highest RMSE value is for 3 categories in 1-4-4 panel design. The highest scores are noted in bold, and the lowest score is noted in bold and italic in Table 5.

Table 5

Findings of RMSE Across All Conditions

Routing Method	Panel Design	3-Category			4-Category			5-Category		
		10	20	30	10	20	30	10	20	30
MFI	1-2	.333	.250	.210	.352	.262	.220	.308	.228	.192
	1-2-2	.330	.247	.208	.350	.260	.219	.307	.227	.191
	1-3	.331	.248	.210	.352	.263	.223	.306	.226	.191
	1-3-3	.327	.244	.209	.350	.261	.220	.304	.224	***.190***
	1-4	.330	.251	.210	.351	.264	.224	.308	.227	.194
	1-4-4	.325	.244	.210	.351	.263	.225	.302	.226	.192
Random	1-2	.381	.292	.244	.382	.288	.241	.335	.252	.210
	1-2-2	.380	.291	.242	**.383**	.289	.243	.336	.252	.210
	1-3	.379	.288	.242	.380	.287	.243	.335	.250	.210
	1-3-3	.378	.288	.244	.380	.288	.245	.333	.253	.211
	1-4	.380	.293	.246	.381	.287	.244	.338	.253	.212
	1-4-4	**.383**	.294	.251	.382	.294	.247	.339	.255	.215

Figure 4

RMSE Values According to Category Numbers

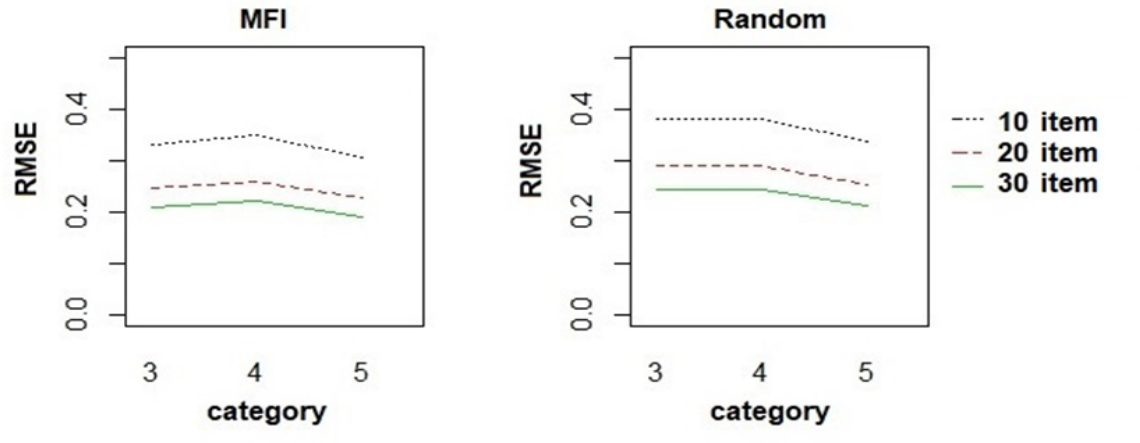


Table 6 shows that ANOVA results for grand mean RMSE indicate that most interaction and main effects were significant. Four factors ANOVA was significant (η^2 =.915). The factors with the highest partial η^2 were the main effects of routing and test length (η^2 =.920). These effects explained 92% of the variance in the mean RMSE. The main effects of test length (η^2=.855) was the factor with the next largest partial η^2. The factor explained about 86% of the variance in the mean RMSE each. Panel design and category added to routing and test length seperatly. The effect was almost the same. When panel design added to routing and test length, the factor explained about 81% of the variance in the mean RMSE (η^2 =.814). When category was added to routing and test length, the factor explained about 81% of the variance in the mean RMSE (η^2 =.810).

Tablo 6

ANOVA Results for Grand Mean RMSE

Factor	Sum of Squares	df	Mean Square	F	p	η^2_p
Routing	2.701	1	2.701	47053.219	.000	.815
Test Lenght	3.630	2	1.815	31625.578	.000	.855
Category	2.002	2	1.001	17440.144	.000	.765
Panel Design	.590	5	.118	2057.510	.000	.490
Routing * Test Lenght	7.088	2	3.544	61744.829	.000	.920
Routing * Category	1.499	2	.750	13060.690	.000	.710
Routing * Panel Design	.196	5	.039	681.434	.000	.242
Test Lenght* Category	.447	4	.112	1945.753	.000	.421
Test Lenght* Panel Design	1.420	10	.142	2474.563	.000	.698
Category * Panel Design	.983	10	.098	1713.316	.000	.616
Routing*Test Lenght * Category	2.620	4	.655	11412.098	.000	.810
Routing * Test Lenght * Panel design	2.685	10	.268	4677.228	.000	.814
Routing * Category *Panel Design	1.614	10	.161	2812.061	.000	.725
Test Lenght* Category * Panel design	1.918	20	.096	1670.814	.000	.758
Routing * Test Lenght * Category * Panel Design	6.645	20	.332	5788.613	.000	.915
Residuals	.614	10692	.000			
Total	864.884	10800				

Correlation

Table 7 shows the correlation values obtained under all research conditions. Figure 5 shows the graphs of correlation values according to the number of categories. Although the MFI routing method generally gives better results than the random routing method, it gives the same results in the 5-category, 20- and 30-item tests. Figure 5 shows the graphs of correlation values according to the number of categories. As the test length increased, the correlation value increased in both routing methods. Similarly, as the number of categories increased, the correlation value increased relatively. The lowest correlation value was found in 1-2, 1-2-2, and 1-4-4 panel designs, in the 4 categories, 10-item test, and in the 1-4-4 panel designs in the 3 category 10-item test and in the random routing method (.923). The highest correlation value was found in all panel designs except 1-4 in the 5-category 30-item test and in MFI (.981). The highest scores are highlighted in bold, and the lowest score is highlighted in bold and italic in Table 7.

Tablo 7

Findings of Correlations Across All Conditions

Routing Method	Panel Design	3-Category			4-Category			5-Category		
		10	20	30	10	20	30	10	20	30
MFI	1-2	.942	.968	.977	.935	.964	.975	.951	.973	**.981**
	1-2-2	.943	.969	.979	.936	.965	.975	.951	.973	**.981**
	1-3	.943	.968	.977	.935	.964	.974	.951	.974	**.981**
	1-3-3	.944	.969	.979	.936	.965	**.975**	.952	.974	**.981**
	1-4	.943	.967	.977	.936	.964	.974	.951	.973	.980
	1-4-4	.945	.970	.979	.936	.964	.974	.953	.973	**.981**
Random	1-2	.924	.956	.969	***.923***	.957	.970	.941	.967	.977
	1-2-2	.924	.956	.970	***.923***	.957	.970	.941	.967	.977
	1-3	.925	.957	.970	.924	.957	.969	.942	.968	.977
	1-3-3	.925	.957	.969	.924	.957	.969	.942	.967	.977
	1-4	.924	.955	.969	.924	.957	.969	.941	.967	.977
	1-4-4	***.923***	.955	.967	***.923***	.955	.968	.940	.966	.976

Figure 5

Correlation Values According to Category Numbers

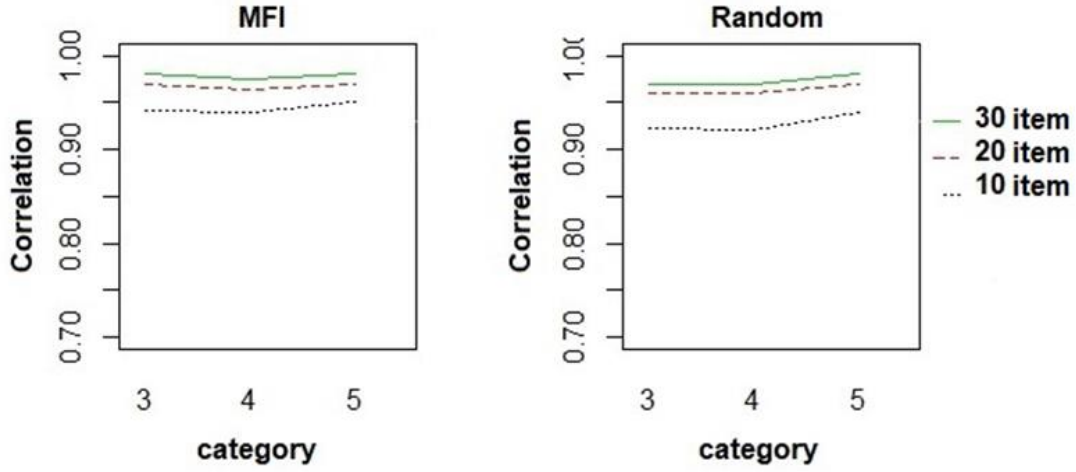


Table 8 shows that ANOVA results for correlation indicate that most interaction and main effects were significant. Four factors ANOVA was significant (η^2 =.893). The factors with the highest partial η^2 were the main effects of routing and test length (η^2 =.898). These effects explained 90% of the variance in correlation. The main effects of test length (η^2=.839) was the factor with the next largest partial η^2. The factor explained about 84% of the variance in the correlation each. When panel design added to routing and test length, the factor explained about 78% of the variance in the correlation (η^2 =.785). When category added to routing and test length, the factor explained about 77% of the variance in the correlation (η^2 =.771).

Tablo 8

ANOVA Results for correlation

Factor	Sum of Squares	df	Mean Square	F	p	η^2_p
Routing	.262	1	.262	38768.600	.000	.784
Test Lenght	.375	2	.188	27778.873	.000	.839
Category	.166	2	.083	12288.887	.000	.697
Panel Design	.057	5	.011	1699.072	.000	.443
Routing * Test Lenght	.634	2	.317	46923.297	.000	.898
Routing * Category	.152	2	.076	11248.700	.000	.678
Routing * Panel Design	.018	5	.004	535.626	.000	.200
Test Lenght* Category	.051	4	.013	1900.051	.000	.415
Test Lenght* Panel Design	.130	10	.013	1928.010	.000	.643
Category * Panel Design	.092	10	.009	1358.231	.000	.560
Routing*Test Lenght * Category	.243	4	.061	9008.331	.000	.771
Routing * Test Lenght * Panel design	.264	10	.026	3902.227	.000	.785
Routing * Category *Panel Design	.141	10	.014	2081.990	.000	.661
Test Lenght* Category * Panel design	.190	20	.009	1405.130	.000	.724
Routing * Test Lenght * Category * Panel Design	.604	20	.030	4473.880	.000	.893
Residuals	.072	10692	.000			
Total	9937.363	10800				

Discussion

Overall, this study investigated the change in the ability estimations of individuals in tests consisting of polytomous items in the computerized multistage test (MST) environment according to the routing methods based on three categories (3, 4, and 5), six-panel designs (1-2, 1-3, 1-4, 1-2-2, 1-3-3, and 1-4-4), three test lengths (10, 20, and 30-item) and two routing methods (MFI and random). The results were then analyzed for mean absolute bias, mean squares of error (RMSE), and correlation values between actual and observed ability levels.

When examining the average absolute bias, RMSE, and correlation values obtained from the item pools consisting of 3, 4, and 5 category items in terms of item categories, the values obtained from the 3 and 4-category item pools are close. Still, the mean absolute bias obtained from the item pool consisting of 4 category items (.23) and RMSE (.29) is the highest. However, the mean absolute bias (.19) and RMSE (.25) values obtained from the item pool consisting of 5-category items are lower than the other categories. In addition, the correlation value (.97) is at the highest level in 5-category items compared to other categories. According to the results obtained, as the number of categories increases, mean absolute bias and RMSE decrease, while correlation values increase.

When examined in terms of routing methods, MFI and random routing methods have similar tendencies, but MFI delivers better results. This was consistent with the previous studies. For example, Macken-Ruiz (2008) compared three routing methods with generalized partial credit model item response theory:

MI, fixed θ, and number-right routing in MST environment, and found that the best performance was observed under the maximum information routing. This was because MFI is a dynamic routing method that calculates module-level information first and, selects the best appropriate module for a test taker. However, the random routing approach, a kind of static method, does not use such an adaptation, and randomly selects the next module among the available modules. This might result in that a test taker with high ability level can receive an easier module at the next level which would inflate his/her ability estimation. Therefore, MFI yielded better results, as also found in Svetina et al. (2019).

Kim et al. (2013) observed that the accuracy of the ability estimates increased as the test length increased. Similarly, in our study mean absolute bias decreased as the test length increased. In addition, as the test length increased, the correlation values also increased. Oztürk (2019) examined how the length and feature of the routing module affect the measurement accuracy in various panel designs. In that study, with two-category items, correlation values increased as the test length increased. As the test length increased, the RMSE value decreased in both routing methods. It can be seen that when examined in terms of test length, the results obtained in our current study show similarities with studies conducted with dichotomous items in the literature (Oztürk, 2019).

Our current study examined an item pool consisting of polytomous items and different conditions, all examined panel designs (1-2, 1-2-2, 1-3, 1-3-3, 1-4, and 1-4-4) showed similar results. Kim et al. (2013) determined all routing methods classification decisions equally well in their studies where they utilized an item pool consisting of polytomous items based on partial credit model (PCM), different panel designs (1-3-3, 1-3-2, 1-2-3, and 1-2-2) and routing methods (ML- DPI, SL-DPI, and M-AMI). Zenisky (2004) did not find meaningful differences between these panel structures or routing methods. The precision of their classification decision was performed all the same. However, some studies have dichotomous items, where the mean error value decreases as we move from the two-stage panel design to the three-stage panel pattern (Sari & Raborn, 2018). Therefore, while the panel design used in MST applications in which an item pool consisting of polytomous items is used does not matter, choosing three-stage panel designs in dichotomous MST applications will provide more accurate results. However, as in the case of Sari and Raborn (2018) and Zenisky (2004), the chosen routing method severely affects the accuracy of the results.

Our study is limited to three kinds of polytomous items (3, 4, and 5 categories), six-panel designs (1-2, 1-3, 1-4, 1-2-2, 1-3-3, and 1-4-4), three test lengths (10, 20, and 30) and two routing methods (MFI and random). According to this study's results, better values were obtained as the number of categories increased. Considering the number of categories in future studies, 5-category items should be preferred. Since there is no difference between the panel designs in the current study, different applications can be made by choosing the panel design suitable for the item pool in future studies. Applications based on actual study parameters can be made with different routing methods (MLWMI, MPWMI, MKL, and MKLP). Classification precision can be examined using different test lengths and item category numbers in MST.

Declarations

Author Contribution: Hasibe Yahsi Sari-Conceptualization, methodology, analysis, writing & editing, visualization. Hülya Kelecioglu-Conceptualization, methodology, writing-review & editing, supervision.

Conflict of Interest: No potential conflict of interest was reported by the authors.

Ethical Approval: Simulated data were used in this study. Therefore, ethical approval is not required.

References

Chen, L-Y. (2010). *An investigation of the optimal test design for multi-stage test using the generalized partial credit model.* [Doctoral dissertation, The University of Texas]. UT Electronic Theses and Dissertations. https://repositories.lib.utexas.edu/handle/2152/ETD-UT-2010-12-344

Choi, Y. J., & Asilkalkan, A. (2019). R packages for item response theory analysis: Descriptions and features. *Measurement: Interdisciplinary Research and Perspectives*, *17*(3), 168-175. https://doi.org/10.1080/15366367.2019.1586404

Dodd, B. G., De Ayala, R. J., & Koch, W. R. (1995). Computerized adaptive testing with polytomous items. *Applied Psychological Measurement, 19*(1), 5-22. https://doi.org/10.1177/014662169501900103

Donoghue, J. R. (1994). An empirical examination of the IRT information of polytomously scored reading items under the generalized partial credit model. *Journal of Educational Measurement*, *31*(4), 295-311. https://doi.org/10.1111/j.1745-3984.1994.tb00448.x

Embretson, S. E., & Reise, S. P. (2013). *Item response theory*. Psychology Press.

Han, K. C. T. (2022). User's Manual: MSTGen. Retrieved from https://www.umass.edu/remp/software/simcata/mstgen/MSTGen_Manual.pdf

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**

ILOG. (2006). ILOG CPLEX 10.0 [User's manual]. Paris, France: ILOG S.A. Retrieved from https://www.lix.polytechnique.fr/~liberti/teaching/xct/cplex/usrcplex.pdf

Kim, J., Chung, H., & Dodd, B. G. (2010, May). *Comparing routing methods in the multistage test based on the partial credit model* [Conference presentation]. In AERA, Denver, CO.

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,* 1087–1098. https://doi.org/10.3758/s13428-013-0316-3

Luecht, R. M. (2000, April). *Implementing the computer-adaptive sequential testing (CAST) framework to mass produce high quality computer-adaptive and mastery tests*. [Conference presentation]. In NCME, New Orleans, LA. https://eric.ed.gov/?id=ED442823

Luecht, R. M., & Nungester, R. J. (1998). Some practical examples of computer-adaptive sequential testing. *Journal of Educational Measurement*, *35*(3), 229-249. https://doi.org/10.1111/j.1745-3984.1998.tb00537.x

Luo, F., Ding, S., Wang, X., & Xiong, J. (2016). Application study on online multistage intelligent adaptive testing for cognitive diagnosis. *Quantitative Psychology Research*, 265-275. https://doi:10.1007/978-3-319-38759-8_20

MacGregor, D., Yen, S. J., & Yu, X. (2022). Using multistage testing to enhance measurement of an english language proficiency test. *Language Assessment Quarterly*, *19*(1), 54-75. https://doi.org/10.1080/15434303.2021.1988953

Macken-Ruiz, C. L. (2008). *A comparison of multi-stage and computerized adaptive tests based on the generalized partial credit model* [Doctoral dissertation, The University of Texas]. ProQuest Dissertations Publishing. https://www.proquest.com/docview/304482829?pq-origsite=gscholar&fromopenview=true

Magis, D., Yan, D., von Davier, A., & Magis, M. D. (2018). Package 'mstR'. Retrieved from https://cran.r-project.org/web/packages/mstR/mstR.pdf

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *ETS Research Report Series, 1992*(1), i–30. https://doi.org/10.1002/j.2333-8504.1992.tb01436.x

Oztürk, N. B. (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

R Core Team. (2018). R: A language and environment for statistical computing: R foundation for statistical computing.

Ridho, A. (2022, January). Sociocultural Literacy Assessment: Validation of Multistage Generalized Partial Credit Testing Design. In *International Conference on Madrasah Reform 2021 (ICMR 2021)* (pp. 382-386). Atlantis Press. https://doi.org/10.2991/assehr.k.220104.056

Rutkowski, L., Liaw, Y. L., Svetina, D., & Rutkowski, D. (2022). Multistage testing in heterogeneous populations: Some design and implementation considerations. *Applied Psychological Measurement*, *46*(6), 494-508. https://doi.org/10.1177/01466216221108123

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, 34* (17). https://psycnet.apa.org/record/1972-04809-001

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

Sari, H.I., Yahsi Sarı, H., & Huggins Manley, A.C. (2016). Computer adaptive multistage testing: Practical issues, challenges and principles. *Journal of Measurement and Evaluation in Education and Psychology*, *7*(2), 388-406. https://doi.org/10.21031/epod.280183

Svetina, D., Liaw, Y. L., Rutkowski, L., & Rutkowski, D. (2019). Routing strategies and optimizing design for multistage testing in ınternational large-scale assessments. *Journal of Educational Measurement*, *56*(1), 192-213. https://doi.org/10.1111/jedm.12206

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.

Weissman, A., Belov, D. I., Armstrong, R. D. (2007). Information-based versus number-correct routing in multistage classification tests. *LSAC Research Report Series*. No. 07-05. Law School Admission Council. https://www.researchgate.net/publication/237288650_Information-Based_Versus_Number-Correct_Routing_in_Multistage_Classification_Tests

Zenisky A., Hambleton R.K.,& Luecht R.M. (2009) Multistage testing: Issues, designs, and research. In: van der Linden W., Glas C. (eds) *Elements of Adaptive Testing.* Springer.

Zenisky, A. L. (2004). *Evaluating the effects of several multi-stage testing design variables on selected psychometric outcomes for certification and licensure assessment* (Publication No. 5710) [Doctoral dissertation, University of Massachusetts Amherst]. UMass Amherst Libraries. https://scholarworks.umass.edu/dissertations_1/5710