ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2023; 14(4); 396-412

Comparison of Different Computerized Adaptive Testing Approaches with Shadow Test Under Different Test Lengths and Ability Estimation Method Conditions

Mahmut Sami YİĞİTER* Nuri DOĞAN**

Abstract

Adaptive testing approaches have been used and adopted in many international large-scale assessments (PISA, TIMSS, PIRLS, etc.). The shadow test approach, on the other hand, is an innovative testing approach that both meets all test specifications and constraints and aims to provide maximum information at the test taker's true ability level. The aim of this study is to investigate the effectiveness of four different adaptive testing approaches created with shadow test (CAT, 2-Stage O-MST, 3-Stage O-MST, and LOFT) according to the test length and ability estimation method. With the Monte Carlo (MC) study in R software, 200 item parameters and 2000 test takers were generated under the 3PL model and the results were calculated over 50 replications. The results show that CAT, 2-Stage O-MST, and 3-Stage O-MST are quite similar in effectiveness, while LOFT is less effective than these techniques. As the test length increases, the measurement precision increases in all different types of adaptive tests. Although the EAP method generally presents better measurement precision than the MLE method, at the extremes of the ability scale, MLE has been found to present good measurement precision. In the research, it is discussed that large-scale assessments can benefit from adaptive testing created with a shadow test approach.

Keywords: computerized adaptive testing, shadow test, on-the-fly multistage testing, linear on-the-fly test

Introduction

Linear tests have been the most popular way of measuring knowledge, skill, and ability in the field of education for centuries. With the advancements in computer hardware and software, Computer Adaptive Testing (CAT) has been adopted and used in many applications worldwide, including the Graduate Record Examination (GRE), Graduate Management Admission Test (GMAT), and Medical College Admission Test (MCAT), as it provides efficient ability estimation and shortens test time (Kirsch & Lennon, 2017; Gökçe & Glas, 2018; Khorramdel et al., 2020; Akhtar et al., 2023; Ebenbeck, 2023).

In linear tests (LT), test takers take all items. A large number of items are needed to obtain effective ability estimation from linear tests (Huang et al., 2009). In CAT, on the other hand, individual tests are obtained by analyzing the properties of the items by algorithms (Raborn & Sari, 2021). CAT is a computer-based test; it can have a fixed or varying length. The test management algorithm presents items to the test taker consecutively and adjusts the difficulty of the items to estimate the test taker's ability level as the test progresses (Wainer, 1990; Hendrickson, 2007; Choi & van der Linden, 2018; Gündeğer & Doğan, 2018). In Computerized Multistage Testing (MST), a group of items called "module" is administered to test takers. In MST, the difficulty of the test is adjusted between modules according to the answers given by the test taker (Yigiter & Dogan, 2023). Although there are studies showing that CAT is more effective than MST and LT in terms of measurement precision (Patsula, 1999; Schnipke & Reese, 1999), MST has beneficial aspects for test administration and test takers (Kim & Plake, 1993; van der Linden, 2010).

* Dr., Social Sciences University of Ankara, Distance Education Application and Research Center, Ankara-Türkiye, e-mail: mahmutsamiyigiter@gmail.com, ORCID ID: 0000-0002-2896-0201
** Prof. Dr., Hacettepe University, Faculty of Education, Ankara-Türkiye, e-mail: nuridogan2004@gmail.com, ORCID ID: 0000-0001-6274-2016

To cite this article:
Yiğiter, M. S., & Doğan, N. (2023). Comparison of different computerized adaptive testing approaches with shadow test under different test lengths and ability estimation method conditions. *Journal of Measurement and Evaluation in Education and Psychology, 14*(4), 396-412. https://doi.org/10.21031/epod.1202599

Received: 11.11.2022
Accepted: 20.10.2023

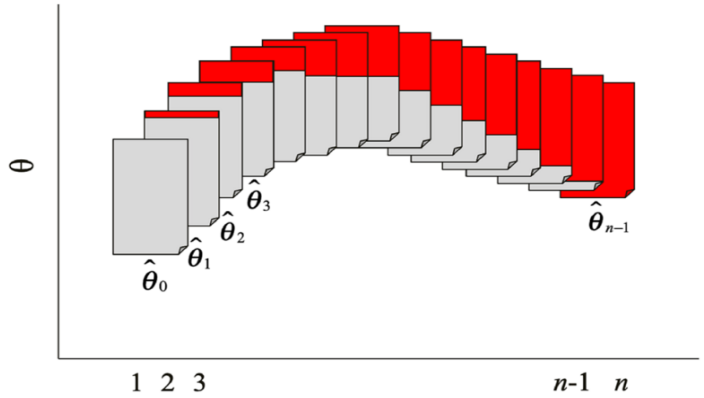
It is important to choose a good model with parameters that reflect the characteristics of the items and the abilities of the test takers in order to maintain the adaptive test successfully. Item Response Theory (IRT) has been successfully operating in its nearly century-old historical development. Many models have been developed under IRT models. In addition to the statistical model, the creation of a well-designed item pool for the attribute to be measured plays an important role in the success of the adaptive test. The selection of optimal items at each ability level, updated with a well-designed item pool and a successful statistical model, can be defined as a mathematical optimization problem. In current adaptive testing algorithms, many methods have been developed as item selection methods, including the maximum Fisher information (MFI), the maximum likelihood weighted information (MLWI) (Veerkamp & Berger, 1997), the maximum posterior weighted information (MPWI) (van der Linden, 1998), the maximum expected information (MEI) (van der Linden, 1998), the minimum expected posterior variance (MEPV), the Kullback-Leibler (KL) divergency criterion (Chang & Ying, 1999), the posterior Kullback-Leibler (KLP) criterion (Chang and Ying, 1996), the global-discrimination index (GDI) (Kaplan et al., 2015). MFI criterion can work successfully in a simple adaptive test where only the item selection from the item pool is based on the amount of information criterion. However, as the number of test specifications and constraints increases in item selection from the item pool, the number of combinations increases rapidly in item selection. Therefore, the complexity of the solution gap in the MFI criterion can make things unsolvable. When the test specifications and constraints such as different contents, item types, word count, expected response time, common stem items, enemy items (items where one item indicates the solution of another item due to the similarity of their content), word count, answer key distribution are considered together, the combinatorial complexity of the problem increases rapidly with each additional constraint. Under such a set of constraints, it is necessary to check each of the possible solutions until the information criterion has the largest value. Also, considering that the test termination rule will end with an incomplete test, there is no guarantee that the test taker will be presented with a test that meets all the constraints (van der Linden, 2022).

The basis of this mathematical optimization problem in adaptive tests is discrete optimization, which requires items to be found in order. Instead of discrete optimization that selects items one by one in each ability estimation, test forms that meet all test specifications and constraints can be created with a mixed integer programming (MIP) methodology that combines a fixed test form. In adaptive testing, Shadow Test Approach has come to the fore with the idea of offering test-takers a test that meets all test specifications and constraints (van der Linden & Chang, 2003).

CAT with Shadow Test Approach

The idea behind the shadow test approach is to create a test that both meets all test constraints and offers maximum information at the test taker's true ability level. The shadow test is obtained by assembling a full-length test that satisfies all the constraints set by the algorithm. When a shadow test is created, the item to be administered to the test taker is the item with the most information. In addition, each shadow test is assembled in such a way that the information function at the interim ability level has the maximum value, and the item to be selected from the shadow test and applied to the test taker has the maximum contribution to this function (van der Linden, 2009). Each subsequent shadow test also includes all items that have already been implemented by the test taker. Therefore, the final shadow test is true adaptive testing and always satisfies all constraints. The basic structure of the shadow test approach is shown in Figure 1.

Figure 1.
Basic Structure of the Shadow Test (van der Linden, 2022)



In Figure 1, the horizontal axis of the graph shows the position of the items in the test; the vertical axis represents the ability level that is updated after the implementation of each item. The higher the vertical position of the shadow tests, the higher the current ability estimation. Towards the end, the convergence of the positions of the shadow tests represents that the final ability estimation has become stable. The red part of the shadow test represents the items answered by the test taker. The gray part represents the part of the ability estimation that is reassembled after a new update (taking into account the items in the red part). The final shadow test includes all of the items actually taken by the test taker (van der Linden, 2009). Different adaptive testing approaches emerged by assembling the freeze-refresh mechanism introduced by van der Linden and Diao (2014) and shadow test at different item locations.

Freeze-Refresh Mechanism and Different Adaptive Testing Approaches

The original shadow test approach is based on reassembling the shadow test at every θ ability update. However, it is stated that instead of a new test assembly after each item, test assembly can be performed at predetermined item locations of the test. With this freeze-refresh mechanism, which was first introduced by van der Linden and Diao (2014), different adaptive testing approaches can be obtained by adjusting the test adaptation points to different item positions. Some adaptive testing approaches that can be obtained by changing the adaptation points according to the locations of the items are shown in Figure 2.

Figure 2.
Different Adaptive Testing Approaches with Freeze-Refresh Mechanism

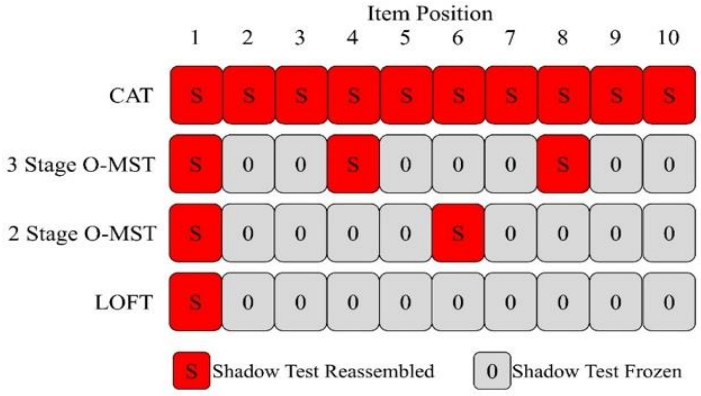


Figure 2 shows that different adaptive testing approaches can be obtained by reassembling the shadow test at different adaptation points on a test consisting of 10 items. Here, the letter "S" in the red box means that the shadow test was reassembled at those item positions and can be referred to as the "adaptation point". In the "0" item positions shown in the gray box, it indicates that the test is frozen (the shadow test is not reassembled) (Choi & van der Linden, 2018). These test approaches, hybrid adaptive tests can be created with the freeze-refresh mechanism. The four different approaches created by the freeze-refresh mechanism and discussed in this study are as follows;

• CAT is a fully adaptive test in which the shadow test is reassembled at each item position;

• 3-Stage O-MST (3 Stage On-The-Fly Multistage Testing) is a three-stage adaptive test in which the shadow test is reassembled at three specified item positions (item positions 1, 4, and 8);

• 2-Stage O-MST (2 Stage On-The-Fly Multistage Testing) is a two-stage adaptive test in which the shadow test is reassembled at two specified item positions (item positions 1 and 6);

• LOFT (Linear On-the-Fly Testing) is a uniquely created fixed test that is brought together at the test taker's initial ability level (Choi & van der Linden, 2018).

Ability Estimation

Many different methods have been developed in the estimation of a test taker's ability in IRT. Frequently used ability estimation methods can be listed as Maximum Likelihood Estimation (MLE) (Birnbaum, 1968), Weighted Likelihood Estimation (WLE) (Warm, 1989), Marginal Maximum Likelihood Estimation (MMLE) (Bock & Aitkin, 1981), Expected a Posteriori (EAP) (Bock & Aitkin, 1981) and Maximum a Posteriori (MAP) (Samejima, 1977; Embretson & Reise, 2000).

The MLE method efforts to get the θ value that maximizes the likelihood function. In cases where the test taker's responses to the items are independent of each other, the product of the response probabilities is defined as the likelihood function, and the point at which this function reaches its maximum is estimated as the ability level. The likelihood function is shown in Equation 1.

$$L(u|\theta) = \prod_{i=1}^{n} P_i(u_i|\theta)^{u_i} * Q_i(u_i|\theta)^{(1-u_i)} \quad (1)$$

In Equation 1, *u* represents the response vector. $P_i(u_i|\theta)$ indicates the probability that the test taker will correctly answer item i at ability level θ. $Q_i(u_i|\theta)$ is equal to 1-$P_i(u_i|\theta)$. n represents the number of items. The value at which this likelihood function is maximum is estimated as the test taker's ability level (θ). Ability level (θ) is solved by iterative methods by taking the derivative of the likelihood function given in the above equation. The most common method used for this purpose is the Newton-Raphson method (Wang & Vispoel, 1998).

The EAP method, on the other hand, is one of the Bayesian ability estimation methods that utilize a priori distributions in ability estimation. Its general formula is shown in Equation 2 (Borgatto et al., 2015):

$$EAP\ (\vartheta) = E(\vartheta|u) = \frac{\int_R \vartheta * L(\vartheta|u) * f(\vartheta) d\vartheta}{\int_R L(\vartheta|u) * f(\vartheta) d\vartheta} \quad (2)$$

In the equation, $f(\theta)$ is the prior distribution function and $L(\theta|u)$ is the likelihood function. In the EAP method, the prior distribution of ability levels must be known. If the a priori distribution is incorrect, the EAP may estimate ability parameters incorrectly (Embretson & Reise, 2000). While the MLE method

is an iterative method, the EAP method is not. Therefore, ability estimations can be obtained faster with EAP.

Choi, Moellering, Li, and van der Linden (2016) compared CAT, O-MST, and a hybrid method combining these two methods using the freeze-refresh mechanism. In this study, the authors generated the parameters of the item pool consisting of 1000 items, assuming that they reflect a real case. The results are quite similar for all three approaches. The researchers concluded that the freeze-fresh mechanism works quite successfully, especially when there are common stem items and test constraints that need to be met, and there is no significant decrease in the measurement accuracy of the test.

Zheng and Chang (2015) compared CAT, O-MST, and F-MST in terms of measurement accuracy. In the study, a real item pool of 352 items from a large-scale assessment was used. The results of this study indicate that CAT and O-MST offer very similar measurement precision and that these two methods offer better measurement precision than F-MST.

van der Linden and Diao (2014) compared five different testing approaches, namely CAT, hybrid CAT, O-MST, F-MST, and LT, with real data sets by simulation. The results of this study show that LT is the least efficient, followed by F-MST. The other approaches, namely CAT, hybrid CAT, and O-MST, are reported to be approximately equally efficient.

Han and Guo (2014) propose an O-MST design that does not include pre-combined test modules and combines a new module on the fly at each stage. The researchers compared the CAT, F-MST, and O-MST designs. The 1-3-3-MST design produced similar measurement accuracy results with the newly developed O-MST design at low iteration shaping, while the new O-MST design produced better measurement accuracy results than F-MST when the number of iterations increased to 100. CAT produces better measurement accuracy than both methods.

Choi and van der Linden (2018) compared the O-MST, which they created using the shadow test approach, with the MST and fixed linear test designs. They concluded that although the measurement accuracy obtained from the O-MST is slightly lower than the CAT, it is better than the fixed linear test.

Purpose of the Study and Research Questions

The purpose of this study is to examine the effectiveness of different adaptive testing approaches created with shadow test according to test length and ability estimation methods. Since the test lengths and ability estimation methods that will be discussed in this study have not been included in any previous study; it is thought that this study will contribute to the development of new approaches. For this purpose, the main research question is:

How does the measurement precision of different adaptive testing approaches change according to different test lengths and different ability estimation methods?

According to the main purpose of the study, the three sub-research questions examined in order to examine this research question in detail are as follows:

1. How does the measurement precision change if different adaptive testing approaches (CAT, 3-Stage O-MST, 2-Stage O-MST, LOFT) are used in the adaptive testing approach?

2. How does the measurement precision change if the different fixed-test lengths (20, 30, 40) are used in the adaptive testing approach?

3. How does the precision of measurement change if different ability estimation methods (MLE, EAP) are used in different adaptive testing approaches?

Method

In this study, Monte Carlo (MC) simulations were performed to compare different adaptive testing approaches (Harwell et al., 1996). MC is a simulation method used to analyze the behavior of statistical

models. In this method, the computer generates data according to probabilistic distributions and allows a comparison of the outputs from the model(s) (Sigal & Chalmers, 2016). In the research, a simulation study was carried out by changing the conditions of four different adaptive testing approaches (CAT, 2-Stage O-MST, 3-Stage O-MST, and LOFT), three different test lengths (20, 30, and 40) and two different ability estimation methods (EAP and MLE). All conditions are crossed with each other. Therefore, in this study, 4x3x2=24 conditions were examined. Analyzes were performed by making 50 replications for each condition.

Data Generation

The R program was used to generate the data and the "TestDesign" package in R was used for the analysis (Choi et al., 2022). Within the scope of the research, the parameters of 200 items were generated based on the 3PL model considering the distributions suggested in the literature (Feinberg & Rubright, 2016; Mooney, 1997; Bulut & Sünbül, 2017). Item discrimination parameters were obtained from $a{\sim}lnN$ (0.2, 0.3) log-normal distribution, item difficulty parameters were obtained from $b{\sim}N$ (0, 1) normal distribution, and item prediction parameters were obtained from $c{\sim}Beta$ (5, 16) beta distribution. Ability parameters were produced from the normal distribution $b{\sim}N$ (0, 1), with 2000 test takers. In addition, assuming that the item pool of the test will consist of three different contents, the item pool is randomly divided into three different content: Content 1 40 items (20%), Content 2 100 items (50%), and Content 3 60 items (30%). Descriptive statistics on item parameters and test taker parameters are shown in Table 1.

Table 1

Descriptive Statistics of Item and Ability Parameters

Parameter	N	Mean	Sd	Min	Max
a	200	1.36	0.24	0.87	1.93
b	200	-0.06	1.07	-2.86	2.81
c	200	0.24	0.08	0.07	0.49
Theta	2000	0.00	1.00	-3.11	2.96

Simulation Conditions

There are conditions that are varied in different adaptive tests created with the shadow test approach. The details of these conditions are explained in the sub-headings below.

Starting Rule

In adaptive testing, the test taker's starting level must be determined before the test can be started. If some information about the test takers is available, it can be used as a starting rule. This information can be students' information (previous course scores, graduation scores, student point average, etc.) or the average of the population (Wang & Vispoel, 1998; Stafford et al., 2019). Since this research was conducted with the simulation and since the population was produced from a normal distribution, the initial ability level was determined as θ=0 for all participants.

Item Selection

Many item selection methods have been developed in adaptive testing approaches with shadow test, especially Maximum Fisher Information (MFI), Maximum Posterior Weighted Information (MPWI), Goal Fisher Information (GFI), Full Bayesian (FB), Empirical Bayes (EB). In this study, the MFI

method, which maximizes the amount of information at the interim ability level, was used (Choi et al., 2022).

Content Balancing

In this study, it was assumed that the item pool consisted of three different contents. Content 1, Content 2, and Content 3 are comprised of 40, 100, and 60 items, respectively (20%, 50%, and 30%, respectively). The contents of the items were determined randomly. The number of items obtained from the contents according to the test lengths is presented in Table 1. In all different adaptive testing approaches, the content distributions given in Table 2 are limited. The number of items that will come from the contents is determined according to their percentages. In Table 2, the content distributions of the adaptive tests according to the test lengths are given.

Table 2

Distribution of the Number of Items to be Obtained from the Contents

Content	Test Length		
	20	30	40
Content 1 (%20)	4	6	8
Content 2 (%50)	10	15	20
Content 3 (%30)	6	9	12

Automated Test Assembly Method

There are many methods used for automated test assembly (glpk, lpSolve, lpsymphony, gurobi, etc.). In this study, the "glpk" (Theussl et al., 2019) method was used for automated test assembly.

Freeze-Refresh Mechanism Item Positions

In computerized adaptive tests created with shadow tests, the item locations where the shadow tests are reassembled and re-presented to the participant student in the freeze-refresh mechanism should be determined. The item locations where the shadow tests are reassembled are given in Table 3.

Table 3

Item Positions in which Shadow Tests Reassembled

Adaptive Test Approach	Test Length		
	20	30	40
CAT	All item position	All item position	All item position
3-Stage O-MST	1., 8. and 14.	1., 11. and 21.	1, 14 and 28.
2-Stage O-MST	1. and 11.	1. and 16.	1. and 21.
LOFT	Only 1.	Only 1.	Only 1.

Ability Estimation Method

In computerized adaptive testing, the method of ability estimation should also be determined. In this study, EAP and MLE methods were used for ability estimation.

Termination Rule

Fixed test length, minimum or maximum test length limit, and standard error threshold methods can be used as termination rules in adaptive test applications with shadow test (Choi et al., 2022). Since the LT and O-MST methods were fixed-length tests in this study, "fixed test length" (20, 30, or 40 depending on the condition) was used for all different adaptive testing as the termination rule.

Data Analysis

In this study, 50 replications were performed for each of the 24 different conditions. The relationships between the true and estimated ability parameters obtained from different adaptive test designs for each condition were interpreted by calculating the Pearson Correlation coefficient, Root Mean Square Error (RMSE), and Mean Absolute Error (MAE) values. The formulas for correlation, RMSE, and MAE values are given below.

$$Correlation = \frac{\sum_{i=1}^{n}(\widehat{\theta_i} - \widehat{\overline{\theta}_i})(\theta_i - \overline{\theta_i})}{(n-1)S_{\widehat{\theta_i}}S_{\theta_i}} \quad (3)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\widehat{\theta_i} - \theta_i)^2}{n}} \quad (4)$$

$$MAE = \frac{\sum_{i=1}^{n}|\widehat{\theta_i} - \theta_i|}{n} \quad (5)$$

The θ_i in the formulas represents the true ability level of the participants, and ($\widehat{\theta_i}$) the estimated ability level. ($\overline{\theta_i}$) and ($\widehat{\overline{\theta}_i}$) denote the mean of the true and estimated ability levels respectively, S_{θ_i} and $S_{\widehat{\theta_i}}$ denote the standard deviation of the true and estimated ability levels respectively, and n denotes the sample size.

Codes were written by the researchers in the R to calculate correlation, RMSE, and MAE values according to the conditions. In addition, RMSE and MAE graphs were created to compare the effectiveness of different adaptive testing approaches according to their ability ranges.

Results

In this section, the findings of the research are given. First of all, the results of the research were examined in a general framework according to all conditions; then the results obtained for each sub-problem of the research were presented under sub-headings. The correlation, RMSE, and MAE values calculated for each of the 24 simulation conditions examined in the study are shown in Table 4.

Table 4

Overall Results from All Conditions

Ability Estimation Method	Test Length	Adaptive Test Type	Correlation	RMSE	MAE
EAP	20	CAT	0.938	0.350	0.275
		3-Stage O-MST	0.935	0.358	0.283
		2-Stage O-MST	0.931	0.367	0.289
		LOFT	0.913	0.410	0.322
	30	CAT	0.956	0.297	0.234
		3-Stage O-MST	0.953	0.305	0.241
		2-Stage O-MST	0.951	0.313	0.246
		LOFT	0.935	0.357	0.277
	40	CAT	0.965	0.266	0.211
		3-Stage O-MST	0.962	0.274	0.216
		2-Stage O-MST	0.961	0.280	0.221
		LOFT	0.947	0.325	0.251
MLE	20	CAT	0.937	0.381	0.297
		3-Stage O-MST	0.932	0.398	0.309
		2-Stage O-MST	0.927	0.409	0.317
		LOFT	0.904	0.450	0.346
	30	CAT	0.955	0.318	0.25
		3-Stage O-MST	0.950	0.336	0.261
		2-Stage O-MST	0.947	0.346	0.268
		LOFT	0.928	0.395	0.301
	40	CAT	0.963	0.283	0.223
		3-Stage O-MST	0.961	0.295	0.229
		2-Stage O-MST	0.958	0.309	0.238
		LOFT	0.940	0.363	0.271

Note. CAT = Computerized Adaptive Testing, O-MST = On-the-fly Computerized Multistage Testing, LOFT = Linear On-The-Fly Test, EAP = Expected a Posteriori, MLE = Maximum Likelihood Estimation.

When Table 4 is examined, it is seen that EAP, one of the ability estimation methods, presents good measurement precision (high correlation and low RMSE-MAE) compared to MLE in all conditions. In all conditions, the measurement precision increases as the test length increases.

CAT provides the best measurement precision in all conditions. While CAT is followed by 3-Stage O-MST and 2-Stage O-MST, respectively, LOFT is seen to be in the last sequence in all conditions. In addition, it can be said that while CAT, 3-Stage O-MST, and 2-Stage O-MST have very similar measurement precision, measurement precision is significantly less because LOFT does not have an adaptation point.

In this section, the correlation, RMSE, and MAE values obtained by averaging from Table 4 and the answers to the three sub-research questions mentioned above were sought. In addition, RMSE and MAE graphs were drawn and interpreted according to their ability ranges.

Results on the First Problem

With regard to the first sub-research question, it was examined how measurement precision changed in different adaptive testing approaches with the shadow test. Findings related to this sub-research question are presented in Table 5.

Table 5

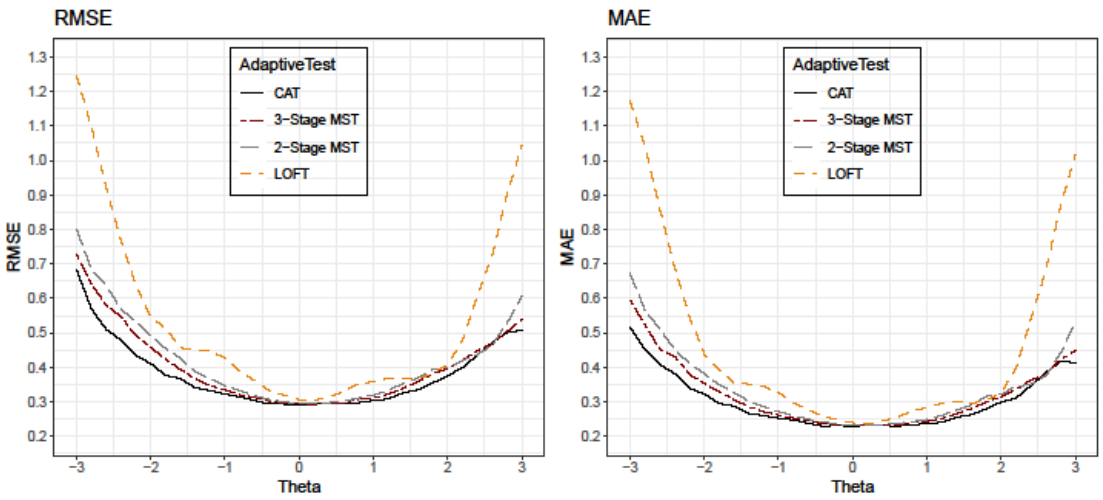
Results by Adaptive Testing Approach Condition

Adaptive Test Type	Correlation	RMSE	MAE
CAT	0.952	0.316	0.248
3-Stage O-MST	0.949	0.328	0.257
2-Stage O-MST	0.946	0.337	0.263
LOFT	0.928	0.383	0.295

As seen in Table 5, CAT shows better measurement precision (high correlation and low RMSE - MAE) than other adaptive tests. In terms of measurement precision, 3-Stage O-MST, 2-Stage O-MST, and LOFT come after CAT. In adaptive tests with the shadow test, it can be said that the measurement precision increases as the adaptation point increases. Figure 3 presents the RMSE and MAE values on the ability scale of different adaptive tests.

Figure 3

Findings on Measurement Precision According to Different Adaptive Testing Approaches



As seen in Figure 3, CAT presents better measurement precision than other adaptive testing approaches across the all ability scale in terms of both RMSE and MAE values. The 3 Stage3-Stage O-MST and 2-Stage O-MST approaches also appear to offer slightly worse but still good measurement precision than CAT. Although LOFT achieves almost as good measurement precision as other adaptive testing approaches around θ = 0 ability level, its measurement precision decreases considerably towards extreme ability levels. Due to LOFT's test assembling at θ = 0 ability level and the absence of an adaptation point, it is seen that the measurement precision of RMSE and MAE values at extreme ability levels decreases significantly compared to other adaptive testing approaches. Although there are not

many individuals at extreme ability levels compared to the normal distribution, it is thought that the measurement precision of LOFT will decrease considerably in skewed or uniform ability distributions.

Results on the Second Problem

With the second sub-research question, it was examined how the measurement precision changed according to the test length of the different adaptive tests with shadow test. Findings related to the second sub-research question are presented in Table 6.

Table 6

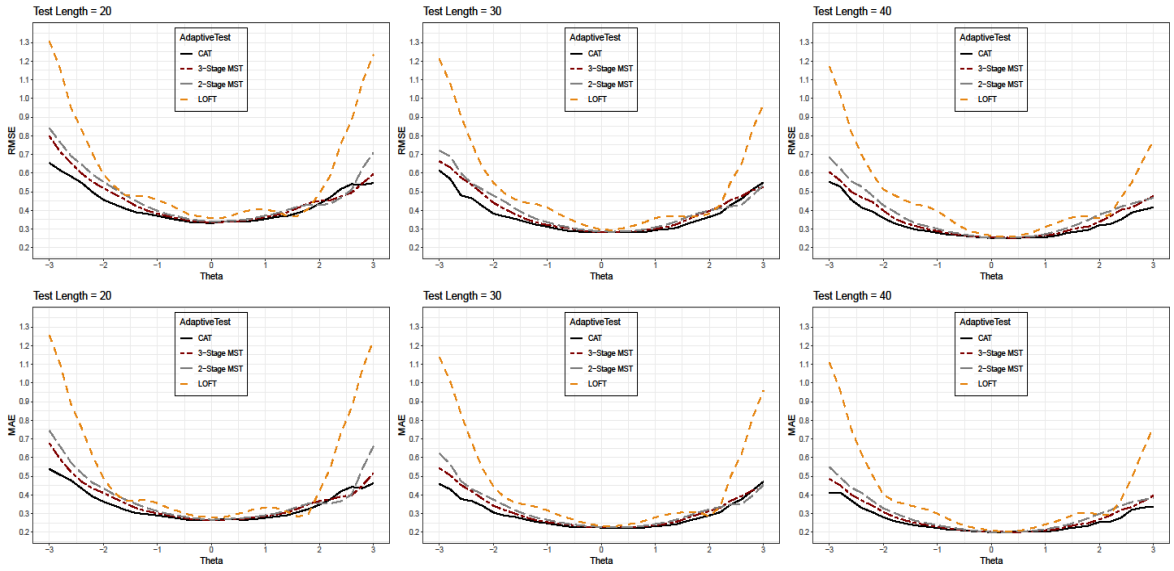
Results by Test Length

N	Correlation	RMSE	MAE
20	0.927	0.390	0.305
30	0.947	0.333	0.260
40	0.957	0.299	0.233

According to the correlation, RMSE, and MAE values, it is seen that the measurement precision increases as the test length increases. The difference in measurement accuracy between 20 and 30 test lengths (ΔCor = 0.020, ΔRMSE = 0.057 and ΔMAE = 0.045) is large, while the measurement precision between 30 and 40 test lengths (ΔCor = 0.010, ΔRMSE = 0.034 and ΔMAE = 0.026) less. This indicates that the measurement precision of 30 to 40 test lengths is more similar than that of 20 test lengths. Figure 4 presents the RMSE and MAE values on the ability scale of different test lengths.

Figure 4

Findings Concerning the Measurement Precision of Test Length by Ability Scale



As seen in Figure 4, both RMSE and MAE values decrease as the test length increases. In addition, as the test length increases, it is seen that the measurement precision at the extreme ability levels decreases more than at the middle ability levels.

Results on the Third Problem

With regard to the third sub-research question, it was examined how the measurement precision changed according to the ability estimation method of the different adaptive testing approaches with shadow test. Findings related to the third sub-research question are presented in Table 7.

Table 7

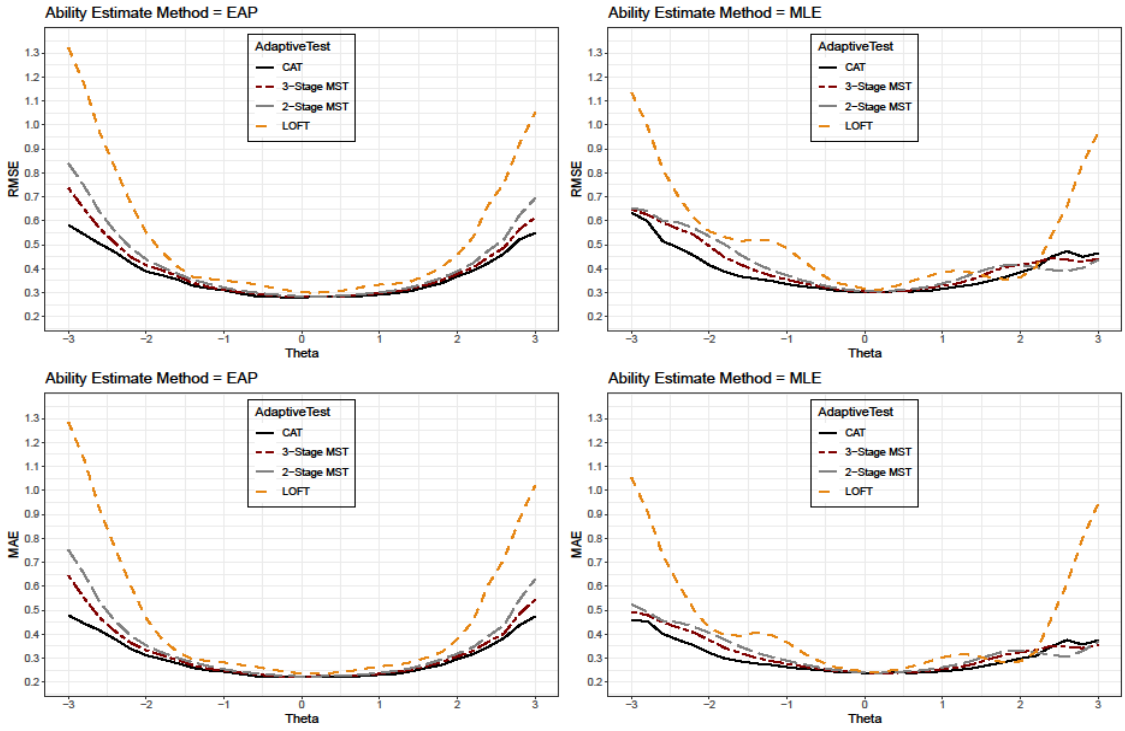
Findings by Ability Estimation Method

Ability Estimate Method	Correlation	RMSE	MAE
EAP	0.946	0.325	0.256
MLE	0.942	0.357	0.276

As seen in Table 7, the EAP method presents better measurement precision than the MLE method according to the correlation, RMSE, and MAE values. In Figure 5, RMSE and MAE values on the ability scale of different ability estimation methods are presented.

Figure 5

Measurement Precision Findings According to Different Ability Estimation Methods in Ability Scale



As seen in Figure 5, the EAP ability estimation method presents better measurement precision than the MLE method in the middle part of the ability scale. On the other hand, it can be said that MLE provides better measurement precision at the extremes of the ability scale. In addition, in terms of both RMSE and MAE, the graphs of the different adaptive testing approaches of the EAP method have a more uniform shape in the ability scale, while the graphs of the MLE method show a more fluctuating increase and decrease.

Discussion

Under the shadow test approach, different adaptive tests can be created with the freeze-refresh mechanism. With this freeze-refresh mechanism, which was first introduced by van der Linden and Diao (2014), adaptive tests such as hybrid-CAT, O-MST, and LOFT can be created since the adaptation points of the test can be adjusted. There are many studies in the literature that shadow tests work successfully under test specifications and constraints that make it difficult for the test algorithm to overcome (van der Linden & Veldkamp, 2004; Choi & Lim, 2022).

This study aims to compare four different adaptive testing approaches created with shadow test according to test length and ability estimation. When the different adaptive test approaches with shadow test are examined in terms of measurement precision, it has been concluded that CAT offers the best measurement precision. It can be stated that 3-Stage O-MST and 2-Stage O-MST follow the CAT, respectively, while LOFT perform worse than other methods in the aspect of measurement precision. Although CAT presents better measurement precision than 3-Stage O-MST and 2-Stage O-MST, it can be said that the RMSE and BIAS values of these three different adaptive testing approaches are quite similar. LOFT, on the other hand, produced worse results than these three approaches because there was no adaptation point.

Choi and van der Linden (2018), in their study on patient-reported outcomes (PRO) measurement, report that CAT offers better measurement accuracy than 3-Stage O-MST and LOFT, similar to the results of this study. In addition, this study states that CAT and 3-Stage O-MST produce very close results. Van der Linden and Diao (2014), in their study comparing different adaptive tests, reported that CAT and O-MST offer very close measurement precision, while fixed-LT offers worse measurement precision than these two adaptive tests. Similarly, Zheng and Chang (2015), in their study comparing CAT, O-MST, and fixed-MST, state that CAT and O-MST offer very similar measurement precision. Comparing the measurement precision with respect to different points of the ability scale, CAT offers very similarly good measurement precision on the 2-Stage O-MST and 3-Stage O-MST ability scales.

Choi et al. (2016), similar to these findings, in shadow tests, it is stated that reassembling the test at each item position and reassembling it at certain item positions will yield nearly equivalent results. On the other hand, LOFT offers good measurement precision in the middle part of the ability scale, similar to other adaptive tests, while measurement precision decreases sharply at extreme ability levels. The reason for the lower measurement precision of the LOFT is the absence of an adaptation point. The results of Han and Guo (2014), van der Linden and Diao (2014), and Choi and van der Linden (2018) are similar to this finding of the study.

In this study, it was concluded that measurement precision increased as the test length increased in different adaptive test types. There are many studies in the literature that adaptive test length increases measurement precision (Weiss, 2004; Özdemir & Gelbal, 2022; Erdem-Kara & Dogan, 2022). Choi and Linden (2018), in their study comparing different adaptive tests with shadow test, states that the 12-item test length offers better measurement precision than 6 items. Xiao and Bulut (2022), in their study examining O-MST, stated that similar to the findings of this study, 60-item length offers better measurement precision than 30 items. The two most important arguments of Computerized Adaptive testing are to reduce test length and increase measurement precision. Therefore, the test length should be short. At the same time, increasing the test length after a certain length will not improve the measurement precision at the desired level due to the "law of diminishing efficiency". In addition, due to the fatigue of the test taker, it may not reflect the real performance of the test taker. In this study, it was concluded that the efficiency of the measurement, which occurs when the test length is increased from 20 to 30, does not occur when it is increased from 30 to 40. It is thought that more research is needed to determine the optimal test length.

It has been concluded that the EAP ability estimation method presents better measurement precision than the MLE method in different adaptive tests with shadow tests. Sahin and Boztunc-Ozturk (2020), in their study on MST, it is seen that EAP performs better than MLE. Similarly, Han (2016) states that EAP performs better than MLE. When the results according to the ability scale are examined, it is seen in the graphs that while the EAP method presents very good measurement precision in the middle area

of the ability scale compared to the MLE in all different adaptive tests, the MLE method presents better measurement precision than EAP in extreme ability scale. The findings of Han (2016) and Şahin and Boztunç-Öztürk (2020) support this conclusion. On the other hand, it can be stated that while the graphs of the ability estimation made with EAP follow a regular path across the all ability scale, the graphs of the ability estimation made with MLE follow a fluctuating path.

Computerized Adaptive Testing and its derivatives have been adopted and used in many large-scale applications over the years. MST has an increasing usage area, especially in recent years. In the PISA (Programme for International Student Assessment) administration implemented in 2018, one MST design was used only in the reading area, out of 3 main areas (Khorramdel et al., 2020). In PISA 2022, it is stated that MST design will be used in more than one area (NCES, 2019). In PIRLS, on the other hand, an MST design consisting of grouped items in 2021 was used (Mullis and Martin, 2019). TIMSS, on the other hand, is prepared to use MST design in both mathematics and science in the 2023 administration (Lin & Foy, 2021). The MST method used in these large-scale assessments is applications where modules and panels consisting of item groups are assembled before the exam administration. Test implementations where modules and panels are assembled before the test application are also called Fixed-MST (F-MST). In F-MST, the adaptation point is low as the test is assembled only according to certain ability levels. Therefore, there are many findings in the literature that the measurement precision of MST is lower than that of CAT (Patsula, 1999; Macken-Ruiz, 2008; Wang, 2017). O-MST, on the other hand, can be considered as a new approach to assembling the advantages of CAT and MST (Zheng and Chang, 2015). As seen in this study, the O-MST approach presents very similar measurement precision to CAT, although the number of adaptations is less. In addition, O-MST has advantages such as presenting items in groups, including items with common stems and passing, skipping, and returning between items, similar to F-MST. In the second and later stages of F-MST, certain ability levels are determined as adaptation points (for example, -1, 0, 1). In O-MST, just like in CAT, every point of the ability level is an adaptation point. This feature of O-MST provides an advantage over F-MST in terms of measurement precision (Han, 2016; van der Linden & Diao, 2014). Given the stated O-MST's advantages, O-MST is promising for international large-scale assessments.

LOFT, on the other hand, presents very similar measurement precision to both CAT and O-MST in the middle area of the ability scale. At extreme ability levels, it differs sharply from both CAT and O-MST by offering rather poor measurement precision. LOFT can be used for diagnostic assessments or to make pass-fail decisions for students with a cut-off point at the midpoint of the ability scale. However, it can be said that its use in exams with cut-off points at extreme ability levels or high-stakes exams will have disadvantages compared to other adaptive tests. In addition, LOFT creates unique linear test forms for each test taker. Therefore, since LOFT does not have any adaptation points, a computer application may not be required. The test forms created with LOFT are applied to the students even in the classroom environment and can be scored after the implementation.

Finally, we offer some practical recommendations. It can be pointed out that a 2 or 3-stage O-MST can be used instead of CAT with some compromise in measurement accuracy. In this way, the advantages of MST can also be utilized. If the scores to be obtained from the test are to be assessed with a cut-off score at extreme ability levels, CAT should be preferred. EAP method can be preferred instead of MAP as an ability estimation method. In terms of test length, shorter tests had lower measurement accuracy, while increasing test length did not linearly increase measurement accuracy. Therefore, it is important to determine the optimal test length in adaptive tests by considering the purpose, content, and measurement accuracy of the test together.

Limitations and Future Studies

This research has some limitations. These limitations can guide researchers in future research. In this study, the fixed test length rule was used as the termination rule. Studies that examine different termination rules can be designed. On the other hand, the MFI method was used as the item selection method in all conditions. The study can be reconsidered with different item selection methods. CAT, O-MST, and LOFT are considered Adaptive Testing Approaches. Hybrid-CAT approaches created by mixing CAT and O-MST can be considered (for more information, see. Choi & van der Linden, 2018). In this study, the item pool was generated by simulation. Working with real item pools is reproducible.

The ability distribution was obtained from the normal distribution. Results can be examined under different or skewed distributions. As an ability estimation method, EAP and MLE were compared. Comparisons can be made with different ability estimation methods such as MAP and MLEF.

Declarations

Author Contribution: Mahmut Sami YİĞİTER: conceptualization, investigation, methodology, data analysis, visualization, writing - review & editing. Nuri DOĞAN: conceptualization, methodology, supervision, writing - review & editing.

Funding: The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Ethical Approval: We declare that all ethical guidelines for authors have been followed by all authors. Ethical approval is not required as the data in this study were generated by a computer program.

Consent to Participate: All authors have given their consent to participate in submitting this manuscript to this journal.

Consent to Publish: Written consent was sought from each author to publish the manuscript.

Competing Interests: No potential conflict of interest was reported by the authors.

References

Akhtar, H., Silfiasari, Vekety, B., & Kovacs, K. (2023). The effect of computerized adaptive testing on motivation and anxiety: A systematic review and meta-analysis. *Assessment, 30*(5), 1379–1390. https://doi.org/10.1177/10731911221100995

Birnbaum, A. L. (1968). Some latent trait models and their use in inferring an examinee's ability. *Statistical theories of mental test scores.*

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443–459. https://doi.org/10.1007/bf02293801

Borgatto, A. F., Azevedo, C., Pinheiro, A., & Andrade, D. (2015). Comparison of ability estimation methods using IRT for tests with different degrees of difficulty. *Communications in Statistics-Simulation and Computation*, *44*(2), 474-488. https://doi.org/10.1080/03610918.2013.781630

Bulut, O., & Sünbül, Ö. (2017). Monte Carlo Simulation Studies in Item Response Theory with the R Programming Language. *Journal of Measurement and Evaluation in Education and Psychology*, *8*(3), 266-287. https://doi.org/10.21031/epod.305821

Chang, H.-H., & Ying, Z. (1999). A-stratified multistage computerized adaptive testing. *Applied Psychological Measurement*, *23*(3), 211–222. https://doi.org/10.1177/01466219922031338

Choi, S. W., & Lim, S. (2022). Adaptive test assembly with a mix of set-based and discrete items. *Behaviormetrika*, *49*(2), 231-254. https://doi.org/10.1007/s41237-021-00148-6

Choi, S. W., & van der Linden, W. J. (2018). Ensuring content validity of patient-reported outcomes: a shadow-test approach to their adaptive measurement. *Quality of Life Research*, *27*(7), 1683-1693. https://doi.org/10.1007/s11136-017-1650-1

Choi, S. W., Lim, S., & van der Linden, W. J. (2022). TestDesign: an optimal test design approach to constructing fixed and adaptive tests in R. *Behaviormetrika*, *49*(2), 191-229. https://doi.org/10.1007/s41237-021-00145-9

Choi, S. W., Moellering, K. T., Li, J., & van der Linden, W. J. (2016). Optimal reassembly of shadow tests in CAT. *Applied psychological measurement*, *40*(7), 469-485. https://doi.org/10.1177/0146621616654597

Çoban, E. (2020). *Bilgisayar temelli bireyselleştirilmiş test yaklaşımlarının Türkiye'deki merkezi dil sınavlarında uygulanabilirliğinin araştırılması*. Yayınlanmamış Doktora Tezi. Ankara Üniversitesi

Demir, S., & Atar, B. (2021). Investigation of classification accuracy, test length and measurement precision at computerized adaptive classification tests. *Journal of Measurement and Evaluation in Education and Psychology*, *12*(1), 15–27. https://doi.org/10.21031/epod.787865

Ebenbeck, N. (2023). Computerized adaptive testing in inclusive education. Universität Regensburg. https://doi.org/10.5283/EPUB.54551

Embretson S. E., & Reise S. P. (2000). *Item Response Theory for Psychologists*. Mahwah, NJ: Lawrence Earlbaum.

Erdem Kara, B., & Doğan, N. (2022). The effect of ratio of items indicating differential item functioning on computer adaptive and multi-stage tests. *International Journal of Assessment Tools in Education, 9*(3), 682–696. https://doi.org/10.21449/ijate.1105769

Feinberg, R. A., & Rubright, J. D. (2016). Conducting simulation studies in psychometrics. *Educational Measurement: Issues and Practice, 35*(2), 36-49.

Gökçe, S., & Glas, C. A. W. (2018). Can TIMSS mathematics assessments be implemented as a computerized adaptive test? *Journal of Measurement and Evaluation in Education and Psychology, 9*(4), 422–436. https://doi.org/10.21031/epod.487351

Gündeğer, C., & Doğan, N. (2018). Bireyselleştirilmiş Bilgisayarlı Sınıflama Testi Kriterlerinin Test Etkililiği ve Ölçme Kesinliği Açısından Karşılaştırılması. *Journal of Measurement and Evaluation in Education and Psychology, 9*(2), 161–177. https://doi.org/10.21031/epod.401077

Han, K. T. (2016). Maximum likelihood score estimation method with fences for short-length tests and computerized adaptive tests. *Applied Psychological Measurement, 40*(4), 289–301. https://doi.org/10.1177/0146621616631317

Han, K. T., & Guo, F. (2014). Multistage testing by shaping modules on the fly. *Computerized multistage testing: Theory and applications*, 119-133.

Harwell, M., Stone, C. A., Hsu, T.-C., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement, 20*(2), 101–125. https://doi.org/10.1177/014662169602000201

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

Huang, Y.-M., Lin, Y.-T., & Cheng, S. C. (2009). An adaptive testing system for supporting versatile educational assessment. *Computers & Education, 52*(1), 53–67. https://doi.org/10.1016/j.compedu.2008.06.007

Kaplan, M., de la Torre, J., & Barrada, J. R. (2015). New item selection methods for cognitive diagnosis computerized adaptive testing. *Applied Psychological Measurement, 39*(3), 167–188. https://doi.org/10.1177/0146621614554650

Khorramdel, L., Pokropek, A., Joo, S. H., Kirsch, I., & Halderman, L. (2020). Examining gender DIF and gender differences in the PISA 2018 reading literacy scale: A partial invariance approach. *Psychological Test and Assessment Modeling, 62*(2), 179-231.

Kim, H., & Plake, B. (1993). *Monte Carlo simulation comparison of two-stage testing and computer adaptive testing*. Unpublished doctoral dissertation, University of Nebraska, Lincoln.

Kirsch, I., & Lennon, M. L. (2017). PIAAC: a new design for a new era. *Large-Scale Assessments in Education, 5*(1), 1-22. https://doi.org/10.1186/s40536-017-0046-6

Macken-Ruiz, C. L. (2008). *A comparison of multi-stage and computerized adaptive tests based on the generalized partial credit model.* Unpublished doctoral dissertation, University of Texas at Austin

Mooney, C. Z. (1997). *Monte carlo simulation*. Sage.

Mullis, I. V., & Martin, M. O. (2019). *PIRLS 2021 Assessment Frameworks*. International Association for the Evaluation of Educational Achievement. Herengracht 487, Amsterdam, 1017 BT, The Netherlands.

National Center for Education Statistics (NCES). (2019). *Program for International Student Assessment 2022 (PISA 2022) Main Study Recruitment and Field Test*.

Özdemir, B., & Gelbal, S. (2022). Measuring language ability of students with compensatory multidimensional CAT: A post-hoc simulation study. *Education and Information Technologies, 27*(5), 6273–6294. https://doi.org/10.1007/s10639-021-10853-0

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

Raborn, A., & Sari, H. (2021). Mixed Adaptive Multistage Testing: A New Approach. Journal of measurement and evaluation in education and psychology, *12*(4), 358–373. https://doi.org/10.21031/epod.871014

Şahin, M. G., & Boztunç Öztürk, N. (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://doi.org/10.21449/ijate.634091

Samejima, F. (1977). A method of estimating item characteristic functions using the maximum likelihood estimate of ability. *Psychometrika*, 42(2), 163-191.

Schnipke, D. L. & Reese, L. M. (1999). A comparison of testlet-based test designs for computerized adaptive testing (Law School Admissions Council Computerized Testing Report 97-01). Newtown, PA: Law School Admission Council.

Sigal, M. J., & Chalmers, R. P. (2016). Play it again: Teaching statistics with Monte Carlo simulation. *Journal of Statistics Education: An International Journal on the Teaching and Learning of Statistics, 24*(3), 136–156. https://doi.org/10.1080/10691898.2016.1246953

Stafford, R. E., Runyon, C. R., Casabianca, J. M., & Dodd, B. G. (2019). Comparing computer adaptive testing stopping rules under the generalized partial-credit model. *Behavior research methods*, 51(3), 1305-1320. https://doi.org/10.3758/s13428-018-1068-x

Theussl, S., Hornik, K., Buchta, C., Schwendinger, F., Schuchardt, H., & Theussl, M. S. (2019). Package 'Rglpk'. GitHub, Inc., San Francisco, CA, USA, Tech. Rep. 0.6-4.

van der Linden WJ, Diao Q (2014). *Using a universal shadow-test assembler with multistage testing*. In: Yan D, von Davier AA, Lewis C (eds) Computerized multistage testing: theory and applications. CRC Press, New York, 101–118

van der Linden, W. J. (1998). Bayesian item selection criteria for adaptive testing. *Psychometrika*, *63*(2), 201–216. https://doi.org/10.1007/bf02294775

van der Linden, W. J. (2009). *Constrained adaptive testing with shadow tests.* Elements of adaptive testing (pp. 31-55). Springer, New York, NY.

van der Linden, W. J. (2010). *Elements of adaptive testing* (Vol. 10, pp. 978-0). C. A. Glas (Ed.). New York, NY: Springer.

van der Linden, W. J. (2022). Review of the shadow-test approach to adaptive testing. *Behaviormetrika*, *49*(2), 169-190. https://doi.org/10.1007/s41237-021-00150-y

van der Linden, W. J., & Chang, H. H. (2003). Implementing content constraints in alpha-stratified adaptive testing using a shadow test approach. *Applied Psychological Measurement, 27*(2), 107-120. https://doi.org/10.1177/0146621602250531

van der Linden, W. J., & Veldkamp, B. P. (2004). Constraining item exposure in computerized adaptive testing with shadow tests. *Journal of Educational and Behavioral Statistics, 29*(3), 273-291. https://doi.org/10.3102/10769986029003273

Veerkamp, W. J. J., & Berger, M. P. F. (1997). Some new item selection criteria for adaptive testing. *Journal of Educational and Behavioral Statistics: A Quarterly Publication Sponsored by the American Educational Research Association and the American Statistical Association*, *22*(2), 203–226. https://doi.org/10.3102/10769986022002203

Wainer, H. (1990). An Adaptive Algebra Test: A Testlet-Based, Hierarchically-Structured Test with Validity-Based Scoring. Technical Report No. 90-92.

Wang, K. (2017). *A fair comparison of the performance of computerized adaptive testing and multistage adaptive testing* (Unpublished Doctoral Dissertation). Michigan State University.

Wang, T., & Vispoel, W. P. (1998). Properties of ability estimation methods in computerized adaptive testing. *Journal of Educational Measurement, 35*(2), 109–135. https://doi.org/10.1111/j.1745-3984.1998.tb00530.x

Wang, T., & Vispoel, W. P. (1998). Properties of ability estimation methods in computerized adaptive testing. *Journal of Educational Measurement*, *35*(2), 109–135. https://doi.org/10.1111/j.1745-3984.1998.tb00530.x

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*(3), 427-450.

Weiss, D. J. (2004). Computerized adaptive testing for effective and efficient measurement in counseling and education. *Measurement and Evaluation in Counseling and Development, 37*(2), 70-84.

Xiao, J., & Bulut, O. (2022). Item Selection with Collaborative Filtering in On-The-Fly Multistage Adaptive Testing. *Applied Psychological Measurement*, 01466216221124089.

Yiğiter, M. S., & Dogan, N. (2023). Computerized multistage testing: Principles, designs and practices with R. *Measurement: Interdisciplinary Research and Perspectives, 21*(4), 254–277. https://doi.org/10.1080/15366367.2022.2158017

Yin, L., & Foy, P. (2021). TIMSS 2023 Assessment Design. TIMSS 2023 Assessment Frameworks, 71.

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