Bartın University Journal of Faculty of Education
BUEFAD Volume 13, Issue 3

547-559
dergipark.org.tr/buefad
DOI: 10.14686/buefad.1199798

Comparison of CAT Procedures at Low Ability Levels: A Simulation Study and Analysis in the Context of Students with Disabilities

Selma Şenel [a*]

a Assoc. Prof. Dr., Balıkesir University, https://orcid.org/0000-0002-5803-0793 *selmahocuk@gmail.com

Research Article
Received: 5.11.2022
Revised: 24.5.2023
Accepted: 29.5.2023

Abstract

The estimation of extreme abilities in computerized adaptive testing (CAT) is more biased and less accurate than that of intermediate abilities. This situation contradicts the structure of CAT, which targets all ability levels. This research aims to determine the procedures that perform better at lower skill levels, in accordance with other ability levels, by comparing the performances of various CAT procedures. In addition, a large-scale test examined whether the determined procedures would show similar performance in the ability levels of students with disabilities, as a group unfortunately more often of extreme abilities and that CAT will offer advantages in many respects. A pool of 1000 items and 1000 examinees with standard normal ability distribution were simulated with Monte Carlo. The CAT performances of 36 conditions consisting of different item selection methods, ability estimation methods and termination rules were compared. As a result of the research, the precision criterion termination rule used together with the maximum likelihood ability estimation method, Kullbak-Leibler information item selection rule, and precision criterion termination rule with test length limit (20 items) performed better and more consistently in terms of CAT performance across the ability levels. These procedures show high performance in the ability levels of students with disabilities, also in real data.

Keywords: Computerized adaptive testing, CAT procedures, extreme ability levels, students with disabilities, Monte Carlo simulation, item selection method, students with low ability.

BOBUT Prosedürlerinin Düşük Yetenek Düzeylerindeki Performanslarının Karşılaştırılması: Simülasyon Çalışması ve Özel Gereksinimli Öğrenciler Bağlamında İnceleme

Öz

Bilgisayar Ortamında Bireye Uyarlanmış Test (BOBUT) yönteminin temel iddialarından biri ölçülen özellik bakımından uçlarda yer alan yeterliklerde geleneksel testlere göre daha kesin ve güvenilir sonuçlar üretmesidir. Ancak, BOBUT'ta da uç yeteneklerin kestiriminin orta yetenektekilere göre daha düşük kesinlikte olduğu, yanlı sonuçlar elde edilebildiği bilinmektedir. Bu durum, BOBUT'un tüm yeterlik düzeylerini hedefleyen yapısına ters düşmektedir. Bu araştırmada, çeşitli BOBUT prosedürlerinin performanslarının karşılaştırarak, alt yetenek düzeylerinde, diğer yetenek düzeyleri ile uyuşan biçimde, daha iyi performans gösteren algoritmaları belirlemek amaçlanmıştır. Ek olarak geniş ölçekli test sonuçlarından yola çıkarak, belirlenen prosedürlerin özel gereksinimli öğrencilerin yeterliklerinde de benzer performans gösterip göstermeyeceği incelenmiştir. Araştırmada öncelikle Monte Carlo simülasyonu ile 1000 maddelik bir madde havuzu ve standart normal dağılım gösteren 1000 kişilik bir yetenek dağılımı oluşturulmuştur. Farklı madde seçme, yetenek kestirimi yöntemleri ve sonlandırma kurallarından oluşan 36 koşulun, uçlarda yer alan bireylerin kestirimindeki BOBUT performansları kıyaslanmıştır. Araştırma sonucunda, En çok olabilirlik yetenek kestirim yöntemi, Kullbak-Leibler bilgisi madde seçme kuralı, standart hata ve madde uzunluğu sınırı (20 madde) ile birlikte kullanılan standart hata test sonlandırma kurallarının; alt yeterlik düzeylerinde en iyi performans göstererek, yeterlik düzeyleri boyunca BOBUT performansı açısından tutarlılık gösteren bir algoritma oluşturduğu gözlenmiştir. Engeli olan öğrencilerin yeterlik düzeylerinde yüksek performans gösterdiği gözlenen ilgili prosedürler, gerçek veri ile onanmıştır.

Anahtar kelimeler: Bilgisayar ortamında bireye uyarlanmış testler, BOBUT prosedürleri, uç yetenek seviyeleri, özel gereksinimli öğrenciler, Monte Carlo simülasyonu, madde seçim yöntemi, düşük yetenekli öğrenciler.

To cite this article in APA Style:

Şenel, S. (2024). Comparison of CAT Procedures at Low Ability Levels: A Simulation Study and Analysis in the Context of Students with Disabilities. *Bartın University Journal of Faculty of Education, 13*(3), 547-559. https://doi.org/10.14686/buefad.119798

INTRODUCTION

Today, one of the most important applications in the discipline of measurement and evaluation is computerized adaptive testing (CAT) (Linacre, 2000; Weiss, 2011). CAT is an application of item response theory (IRT), the fundamental up-to-date measurement theory. In simple terms, CAT is a method for estimating ability levels with high precision by directing the items closest to the ability level of the respondent. For example, suppose a respondent is extremely poor in terms of the feature that a test measures. When the examinee encounters moderate or hard items, they will not be able to answer any item correctly. In this case, we cannot obtain information about the examinee's abilities. However, if they encounter items that are close to their ability, in other words, easier items, they will probably be able to answer some of them correctly. In this case, we will obtain more information about the examinees' competencies and inadequacies. In addition, the duration of the completed test will be 50% to 80% shorter (Kezer & Koç, 2014; Şenel & Şenel, 2018; Wainer et al., 2000a) since the items that will not provide information about respondents will not be applied. CAT achieves these powerful features thanks to the CAT algorithm; a fundamental example is presented in Figure 1.

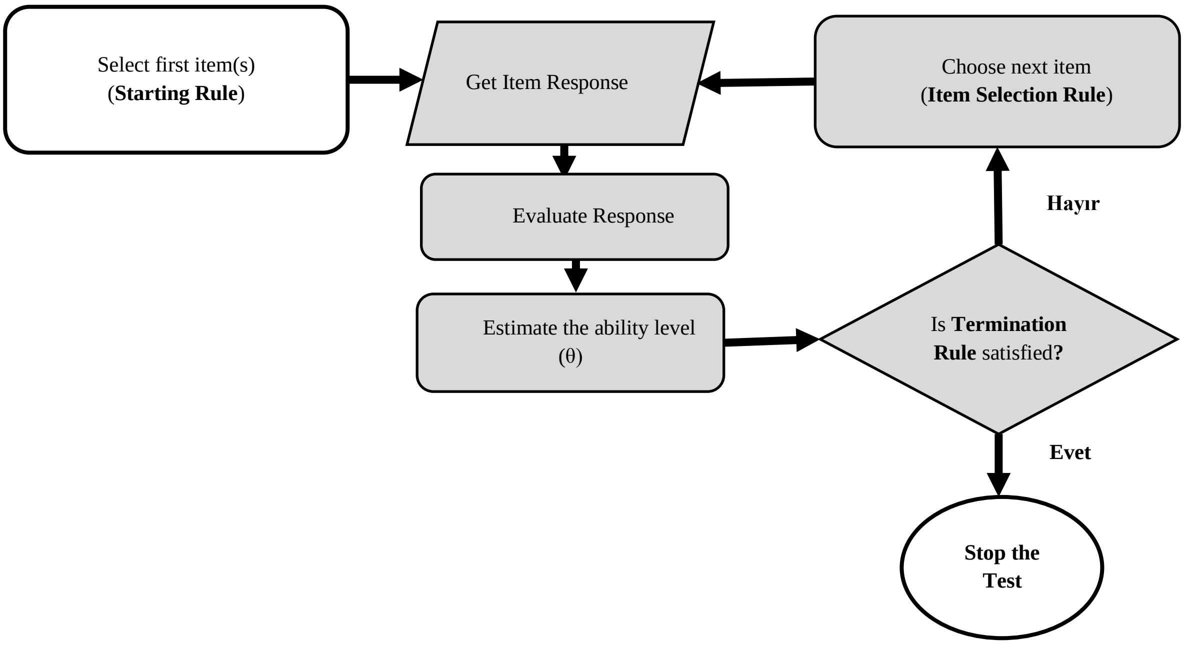

Figure 1. *A Fundamental CAT Algorithm (Şenel, 2021)*

As seen in the CAT algorithm in Figure 1, how the first and next items will be selected and how the test will be terminated are predetermined. The first item is selected from a large item pool consisting of qualified items with IRT-calibrated and predicted psychometric properties. For the first application, the selection method is often applied from items that are of moderate difficulty (Segall, 2004) or that address the medium ability level (θ=0) (Magis et al., 2018). After applying the first item(s), an ability estimation will be predicted. Various methods are used in ability estimation, such as maximum likelihood (ML), maximum a posteriori (MAP), Bayesian expected a posteriori (EAP), weighted likelihood (WL) and robust estimator (RE) (Embretson & Reise, 2000; Magis et al., 2018; Mislevy & Bock, 1982; Segall, 2004; Warm, 1989). The ML (Lord, 1980) method is the most popular ability estimation method and the second most popular method is EAP (Bock & Mislevy, 1982).

After the examinee's first θ estimation, the items are mainly selected from among the ones that will provide high information and are closest to the estimated θ (Şahin & Ozbasi, 2017). Wainer et al. (2000) considered the *item selection rule* as one of the three basic dimensions that affect the validity of CAT. The process to "administer the appropriate item to the examinee" based on CAT takes place in this step. The maximum Fisher information (MFI) criterion, b optimal (*bOpt*), maximum likelihood weighted information (MLWI) criterion, maximum posterior weighted information (MPWI) criterion, Kullback-Leibler (KL) divergence criterion, and θ optimal (*thOpt)* are among the most frequently used item selection methods (Barrada et al., 2009; Magis et al., 2018; van der Linden et al., 2006). Using one of the preferred item selection methods, the most appropriate item for the respondent at that stage of the test is applied. Re-estimation of θ is made after each item response. The "θ estimation-item selection-θ estimation" cycle continues until the termination rule is satisfied.

Various termination rules determine after the administration of which items test will be terminated. The most frequently used termination method is the *precision criterion (PC)* (van der Linden & Glas, 2010). In this method, the test ends when the standard error of ability estimation falls below a certain criterion (Embretson & Reise, 2000), frequently 0.32. The fact that the standard error has decreased to a certain level indicates that the reliability of the result has reached an acceptable level. In addition, a test length limit is an approach applied to terminate the test (Babcock & Weiss, 2009). Another approach is to terminate the test when no item in the item pool provides a predetermined level of information (Maurelli & Weiss, 1981). Using different termination rules together is also a recommended approach (Babcock & Weiss, 2009).

With these features, CAT provides test applications that are consonant with the ability level of the respondent and produces highly reliable test results with each respondent taking a different number of items. In this way, it produces more accurate and reliable test results than traditional tests for extremely low and high ability levels. This feature is one of the main strengths of CAT. With this strength, CAT is preferable for individuals with extreme ability levels. Considering that individuals with special needs such as students with disabilities remain at lower ability levels (Stone & Davey, 2011), CAT is an important option to increase test validity. In addition, CAT is becoming increasingly common in the field of health diagnosis (Gibbons et al., 2014, 2016), and disease is extreme values in health-related measurements. CAT has additional advantages for students with disabilities. Being convenient for computer-based test accommodations is one of them. Apart from this, there is no need for extended time-test accommodations with relatively short tests. Additionally, CAT is preferred because it provides more information and more reliable test scores for students with disabilities (Şenel & Kutlu, 2018a, 2018b; Stone & Davey, 2011).

The estimation precision of CAT is higher than that of conventional tests. However, CAT is less accurate in estimating extreme abilities than intermediate abilities, and biased results are obtained in extreme abilities (Babcock & Weiss, 2009; Riley et al., 2007). There may be positive bias in low ability levels and negative bias in high ability levels. CAT produces more precise estimates near θ=1 (Magis et al., 2018). This situation contradicts the structure of CAT, which targets all ability levels. That CAT produces more biased test results at lower ability levels compared to average ability levels is problematic.

The first reason that comes to mind for the differentiation of estimation precision at different ability levels may be insufficient qualified items in the item pool for extreme abilities (Belov & Armstrong, 2009). The main determinant of CAT performance is the quality of the item pool and its suitability for the target ability (van der Linden et al., 2006). High CAT performance requires an item pool of psychometrically strong items that address a broad θ level (Weiss, 1973). CAT item pools commonly include more items with moderate difficulty and provide more information for average θ level respondents. Figure 2 shows two indicator charts of a dichotomous two-parameter logistic model CAT item bank by Magis et al. (2018).

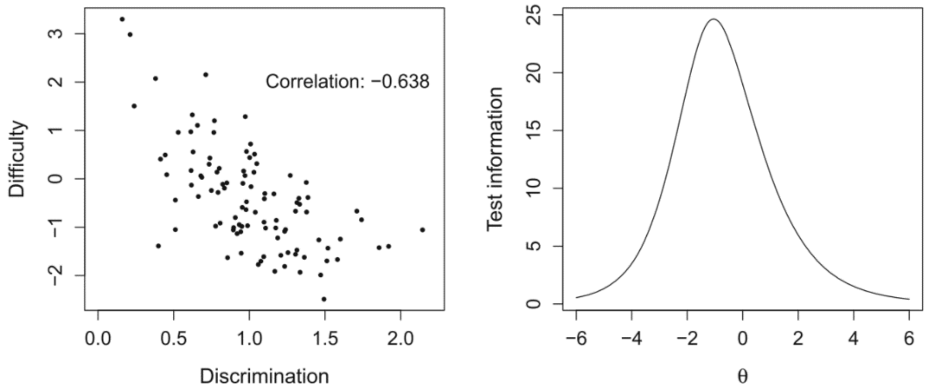

Figure 2. *Scatterplot of Discrimination and Difficulty Coefficients and Test Information Function of the 2PL Item Bank (Magis vd., 2018)*

As can be seen in the example in the figure, the difficulty and discrimination parameters of the items are stacked at moderate levels. Difficult or easy items are fewer in number than items with moderate difficulties.

Regarding this situation, the test mostly provides information at average levels, near θ=1. The information provided by the test decreases at extremely high and low θ levels.

There are studies on how to optimize the item pool and how to apply the most optimal CAT to the targeted population (Belov & Armstrong, 2009; Reckase, 2010). Apart from the design of the item pool, the methods preferred in the CAT algorithm may also affect the performance of CAT for individuals at extreme ability levels. Although the item pool is the main determinant, the methods applied can also affect the optimal CAT. A wide variety of CAT procedure combinations can be created by combining different test entry rules, item selection methods, ability estimation methods, and test termination rules. As a result of each combination, the individual may encounter item x versus item y. In different CAT procedures, approximate but different θ values are produced. As a result, the performance of different algorithms at extreme ability levels is also a critical issue. However, no research has been found in the literature examining which methods are more powerful according to the θ range. For a similar purpose, only concerning ability estimation methods, Chen, Hou, Fitzpatrick ve Dodd (1997) investigated the effect of population distribution and method of θ estimation on CAT using the rating scale model. Consequently, for either normal or negatively skewed population distributions, the three methods of ML estimation, EAP with a normal prior, and EAP with a uniform prior performed similarly. In addition, most studies compare the general performances of CAT. He, Diao ve Hauser (2013) compared the weighted deviation model, weighted penalty model, maximum priority index and shadow test approach item selection procedures in severely constrained CAT. The results indicate that, among all candidate methods, the shadow test approach works the best in terms of measurement accuracy and constraint management, except that it makes the poorest use of items. Some studies also examine CAT procedures that are effective in multidimensional computerized adaptive testing (MCAT) (Seo & Weiss, 2015; Yao, 2013). According to Yao (2013), the optimal five procedures are minimum angle, volume, minimizing the error variance of the linear combination, minimizing the error variance of the composite score with the optimized weight, and KL information. According to Seo and Weiss (2015), the MCAT model without a guessing parameter functioned better than the MCAT model with a guessing parameter. The MAP estimation method provided more accurate θ estimates than the EAP method under most conditions, and MAP showed lower observed standard errors than EAP under most conditions, except for a general factor condition using D_s-optimality item selection.

Common indicators of CAT performance are *test lengths, standard error values, bias,* and *root-mean-square-error(RMSE) values*. CAT with fewer items, low standard error values, low RMSE values, and close to zero bias performs well. Equations for bias and RMSE values are presented in Equation 1 and Equation 2, where j represents the number of respondents and N is the total number of respondents. According to Equations 1 and 2, the high difference between the estimated ability level and the actual ability level proves the low CAT performance.

$$Bias = \frac{\sum_{j=1}^{N} (\widehat{\theta_J}_j - \theta_j)}{\mathrm{N}} \qquad \text{(Equation 1)}$$

$$RMSE\ (Root\ Mean\ Squared\ Error) = \sqrt{\frac{\sum_{j=1}^{N} (\widehat{\theta_J}_j - \theta_j)^2}{\mathrm{N}}} \qquad \text{(Equation 2)}$$

The deduction is that methods that do not show significant changes in the precision of estimations according to the skill range should be selected in CAT applications, which are applied to individuals from a wide range of abilities. It is important to choose algorithms that provide sufficient information about lower ability levels, especially in large-scale tests involving many students with special needs. Based on this determination, this study aimed to compare the performances of various CAT algorithms at lower skill levels. In addition, it examined whether students with special needs who take a large-scale test are at lower proficiency levels than in the literature. The ability levels of students with special needs of the optimal CAT procedure, which emerged from the research, will be examined. In this context, we created the following research questions:

- Which combination of item selection method, ability estimation method, and termination rule has consistent and high CAT performance across different ability levels?
- Is the reading comprehension ability of students with disabilities who took a large-scale test significantly lower than that of students without disabilities?
- How does the optimal CAT algorithm perform at the ability levels of
    - a mixed group of students with and without disabilities who took a large-scale test?

- students with disabilities who took a large-scale test?
- students without disabilities who took a large-scale test?

Definitions

Central exam: The Turkey Central Secondary Education Exam (Ministry of National Education, 2018) for transition to secondary education applied in Turkey.

Reading comprehension ability: The test point obtained in the Turkish subtest of the central exam.

Limitations

In the study, the classification of disabled individuals was based on the disabled student classification based on the Central Exam application.

Post-hoc simulation in the study is limited to the 20-item Turkish subtest of the Central Exam.

METHOD

Study Group

This study has three main research questions. For the first research question, a Monte Carlo simulation was performed. Simulation data and related details are presented in the Data and Data Collection section.

For the second and third research questions, the Turkey Central Secondary Education Exam (hereinafter referred to as the *central exam*) data were analyzed as large-scale test data. Thus, the study group consists of 8th-grade students with disabilities (n=4410) who participated in the central exam in the 2017-2018 academic year and 5000 secondary school students without disabilities randomly selected from the students who participated in the central exam. The study group thus consists of 9410 students in total. The non-disabled group, who did not receive any test accommodation in the exam, constitutes 53.1% of the study group. The students with disabilities (n=4410), were classified into 11 groups. Group information and whether they received extended time accommodation are presented in Table 1. The disability classification is based on the classification used in the central exam.

Table 1. Disability and Extended Time Accommodation Distribution of Study Group.

		extended time accommodation			
Disability group		Yes	No	Total	
Physically impaired	n	253	162	415	9.4%
	%	61.0%	39.0%	100.0%	
Homeschooling and taking test at home	n	11	82	93	2.1%
	%	11.8%	88.2%	100.0%	
Visually impaired	n	355	0	355	8.0%
	%	100.0%	0.0%	100.0%	
Taking test at home	n	1	5	6	0.1%
	%	16.7%	83.3%	100.0%	
Attention deficit and hyperactivity	n	334	0	334	7.6%
	%	100.0%	0.0%	100.0%	
Hearing impaired	n	388	21	409	9.3%
	%	94.9%	5.1%	100.0%	
Mentally impaired	n	1164	85	1249	28.3%
	%	93.2%	6.8%	100.0%	
Pervasive developmental disorder	n	118	0	118	2.7%
	%	100.0%	0.0%	100.0%	
Specific learning difficulty	n	998	0	998	22.6%
	%	100.0%	0.0%	100.0%	
Chronic disease	n	0	59	59	1.3%
	%	0.0%	100.0%	100.0%	
Multiple disabilities	n	362	12	374	8.5%
	%	96.8%	3.2%	100.0%	
Total	n	3984	426	4410	100.0%
	%	90.3%	9.7%	100.0%	

Table 1 provides a summary of the students with disabilities as part of the study group. According to the table, 90.3% of the disabled part of the study group took the test with extended time accommodation. The highest rate among all disability groups is those with mental disabilities (28.3%) and those with specific learning difficulties (22.6%). The cross-table of the study group according to gender and educational institutions is presented in Table 2.

Table 2. Gender and school Type Distribution of Study Group

			School Type				Total	
			Public School	Private School	Religious School	Boarding School	n	%
Gender	Female	n	3496	289	460	45	4290	45.59%
		%	81.49%	6.74%	10.72%	1.05%	100.00%	
	Male	n	4247	369	452	52	5120	54.41%
		%	82.95%	7.21%	8.83%	1.02%	100.00%	
Total		n	7743	658	912	97	9410	
		%	82.28%	6.99%	9.69%	1.03%	100.00%	

According to Table 2, 45.59% of the group are female and 54.41% are male. Over four-fifths of the group (n=7743, 82.28%) are educated in public schools. This situation shows a distribution that reflects the school distribution in Turkey.

Data and Data Collection

Simulation study data

In the research, an item pool of 1000 items was created with ideal item parameter distributions (parameter a, uniformly distributed, in the range of 1-2; parameter b, uniformly distributed, in the range of -3-+3; parameter c, uniformly distributed, in the range of 0-0.20) (Wainer et al., 2000b), with Monte Carlo simulation. An ability distribution of 1000, with a mean of "0", and a standard deviation of "1" with a normal distribution was simulated. The CAT performances of different item selection methods, ability estimation methods and termination rules in estimating respondents at the extremes were compared.

The CAT procedures to be compared were chosen among the frequently preferred and recommended methods in the literature. EAP, ML estimation, and MAP were preferred as ability estimation methods. As item selection rules, MFI criterion, bOpt, KL divergence criterion, thOpt, proportional and progressive methods were used. The *PC* termination rule has been called the most powerful method for estimating low ability levels (Babcock & Weiss, 2009; Choi et al., 2011). The use of the test length limit as a termination rule is not recommended at lower ability levels. However, without item exposure control or content balancing, the ideal length can be specified as 15-20 in the test of a one-dimensional construct (Babcock & Weiss, 2009). In this study, the item length of "20" was added to the conditions as a second termination criterion (PC + 20). A total of 36 conditions were compared using combinations of these methods.

Large-scale test data

Large-scale test data were used to answer the second and third research questions of the study. The data of the students who participated in the 2018 Central Exam (Ministry of National Education, 2018) were obtained with the permission of the Ministry of National Education. Consisting of 90 multiple choice items, the central exam has verbal and numerical parts, which are administered in two separate sessions. The verbal part consists of Turkish, Religious Culture and Morals, History of the Republic of Turkey and Kemalism, and Foreign Language subtests. The numerical part consists of Mathematics and Science subtests. In the research, since the analyses were carried out according to the one-dimensional IRT, a single subtest was studied. Considering that reading comprehension is a basic skill, the Turkish subtest consisting of 20 items was chosen for analysis.

Turkish subtest

To use one-dimensional IRT models, assumptions are needed. The assumptions of unidimensionality, local independence, and model-data fit (Embretson & Reise, 2000; Hambleton & Swaminathan, 1985) were examined before data analysis. To examine the unidimensionality of the Turkish test, a modified parallel analysis was performed and the scree-plot graph in Figure 3 was created.

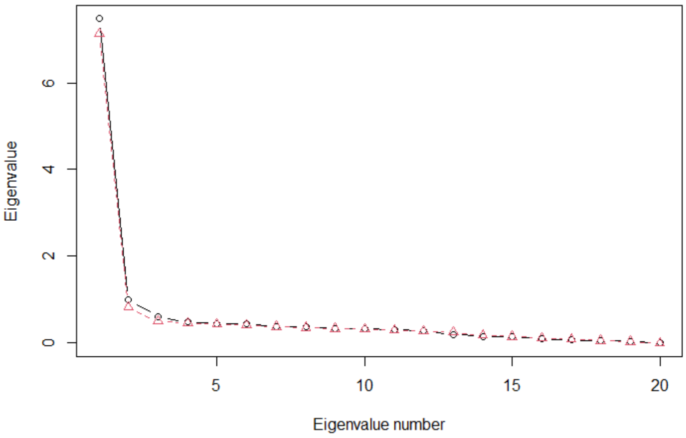

Figure 3. *Scree-plot for Turkish Subtest*

According to Figure 3, the test measures a dominant dimension. To determine which IRT model the test is compatible with, Akaike Information Criterion (ACI) Bayesian Information Criterion (BIC) and log likelihood values were examined according to a one-parameter logistic model, a two-parameter logistic model, and a three-parameter logistic model (3PL), and whether there was a significant difference between model fits was analyzed. The results are presented in Table 3.

Table 3. Model Selection Values

IRT Model	AIC	BIC	Log likelihood	P
1PL	216741.7	216891.9	-108349.9	<0.001
2PL	212752.5	213038.5	-106336.3	
3PL	210676.2	211105.1	-105278.1	

According to Table 3, the model with the highest model fit is the 3PL model. Item parameters were calculated according to the compatible 3PL model and θ values were estimated. In Figure 4, item characteristic curves and test information functions of 20 items in the subtest are presented.

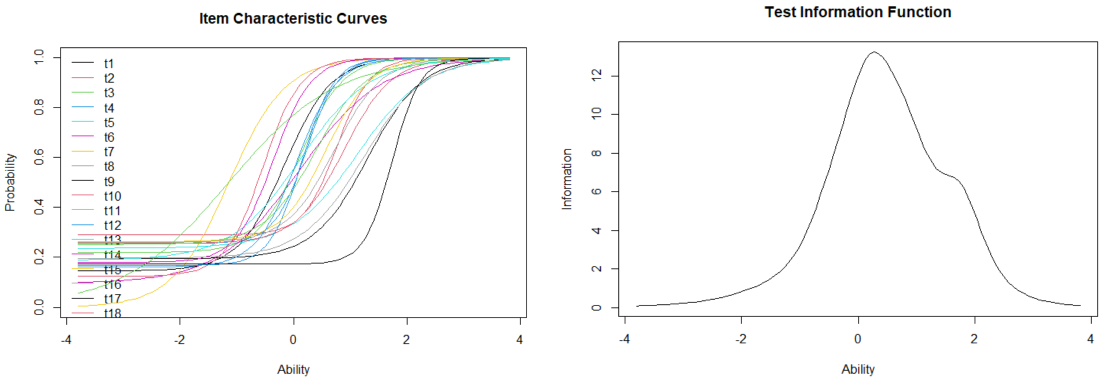

Figure 4. *Item Characteristic Curves and Test Information Functions of Turkish Subtest*

Data Analysis

According to the item characteristic curves in Figure 4, there are items with various difficulties (b parameter) and discriminations (a parameter) in the test. It is observed that the c parameters are around 0.20 as in most tests in multiple choice items. According to the test information function, the test gives the most information is around "0-0.5" θ levels.

θ levels representing students' reading comprehension ability levels were obtained from the Turkish subtest. These θ levels were used to answer the second and third research questions. An independent sample t-test was conducted to examine whether the reading comprehension proficiency of students with disabilities who took the test was significantly lower than that of students without disabilities. It was observed that the data provided the assumption of normality.

Since the real large-scale test data consisted of 20 items, a CAT simulation could not be carried out by considering the applied items as an item pool. A larger item pool, such as 200 or more items, is recommended for CAT (Şahin & Weiss, 2015). Therefore, a 1000-item item pool can be combined with Monte Carlo simulation, with item parameter distributions that can be considered ideal (parameter a is uniformly distributed, in the range of 1-2; parameter b is uniformly distributed, in the range of -3-+3; parameter c is uniformly distributed, in the range of 0-0.20) (Wainer et al., 2000b). The CAT simulation was continued with this simulated item pool. The item selection rule, ability estimation method and termination rule found in the answer to the first research question (item selection rule: KL, ML for ability estimation and PC [SE=0.32] as termination rule) were used in the CAT simulation in the third research question.

Based on the θ levels, CAT simulations were conducted for the third research question. For the ability distribution of the entire study group, CAT simulations were carried out separately based on the ability distributions of students with disabilities and without disabilities. Simulation analyses were carried out in the R *catR* package. Real-data θ estimations and IRT assumptions analysis were carried out in the R *ltm* package.

Research Ethics

The actual data used in the research were obtained from the Ministry of National Education of Turkey with permission. Since the data does not contain personal information, the research complies with ethical principles.

FINDINGS

Which combination of item selection method, ability estimation method, and termination rule has consistent and high CAT performance across different ability levels?

As a result of the research, 11 CAT procedure conditions were determined for optimal performance (r [correlation between actual θ and estimated θ] >= 0.95; bias <=0.01; RMSE< 0.33; number of items < 18) in terms of average CAT performance indicators. These optimal 11 conditions and CAT performance indicators are presented in Table 4.

Table 4. 11 Optimal Performing CAT Conditions

Condition No	Ability Estimation Method	Item Selection Rule	Termination Rule	Average Test Length	r	RMSEA	Bias
1	EAP	MFI	SH	12,9	0,96	0,3032	0,0074
2	EAP	progressive	SH	14,8	0,96	0,295	0,0014
3	EAP	proportional	SH	17,8	0,95	0,3229	0,0034
4	EAP	thOpt	SH	21,9	0,95	0,3287	-0,0063
5	EAP	bOpt	SH	22	0,95	0,3268	-0,0123
6	EAP	KL	SH	13,3	0,95	0,3161	-0,0053
12	ML	KL	SH	14,6	0,96	0,3172	0,0112
18	MAP	KL	SH	12,5	0,95	0,3242	0,0257
19	EAP	MFI	SH + 20	12,9	0,95	0,3101	-0,0053
20	EAP	progressive	SH + 20	15	0,95	0,3261	0,0047
21	EAP	proportional	SH + 20	17,6	0,95	0,3112	-0,0015
24	EAP	KL	SH + 20	13,3	0,95	0,318	-0,0064
30	ML	KL	SH + 20	14,6	0,96	0,3156	-0,0012

According to Table 4, the EAP ability estimation method and the KL divergence item selection criterion outperform other methods in terms of the overall average. To observe the strength of these 11 prominent conditions at different ability levels, the change graph of the RMSEA value according to the θ intervals is presented in Figure 5, and the variation graph of the bias value according to the θ intervals is presented in Figure 6. The numbers for the conditions were presented in Table 1.

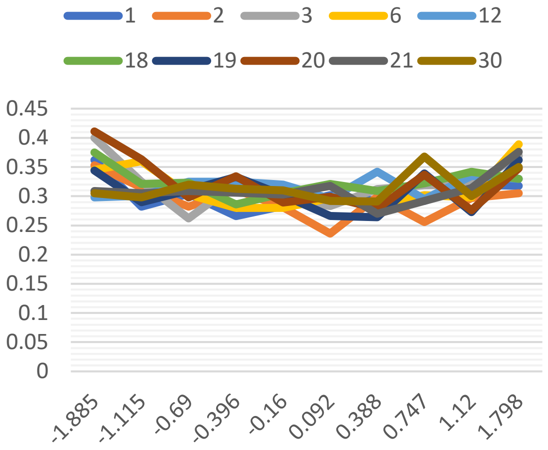

Figure 5. *RMSEA Change at θ intervals*

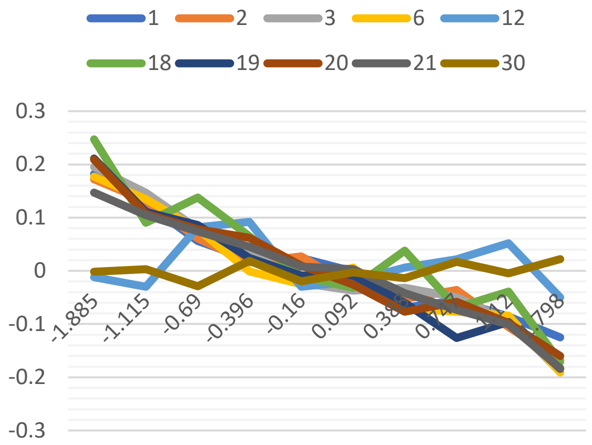

Figure 6. *Bias Change at θ Intervals*

According to Figures 5 and 6, among the algorithms with low mean of RMSEA and bias, the algorithms that show consistency in terms of CAT performance across different proficiency levels are more straight-line, with the best performance at the lower proficiency levels. These can be specified as ML ability estimation method, KL item selection method, PC (condition 12) and PC +20 (condition number 30) test termination rules. It is observed in the chart that the lines of these conditions maintain their low levels.

Is the reading comprehension ability of students with disabilities who took a large-scale test significantly lower than students without disabilities?

This research question tested whether students with special needs stated in the literature mostly had low proficiency levels. First, Figure 7 presents histogram graphs of the estimated θ levels of students with and without disabilities.

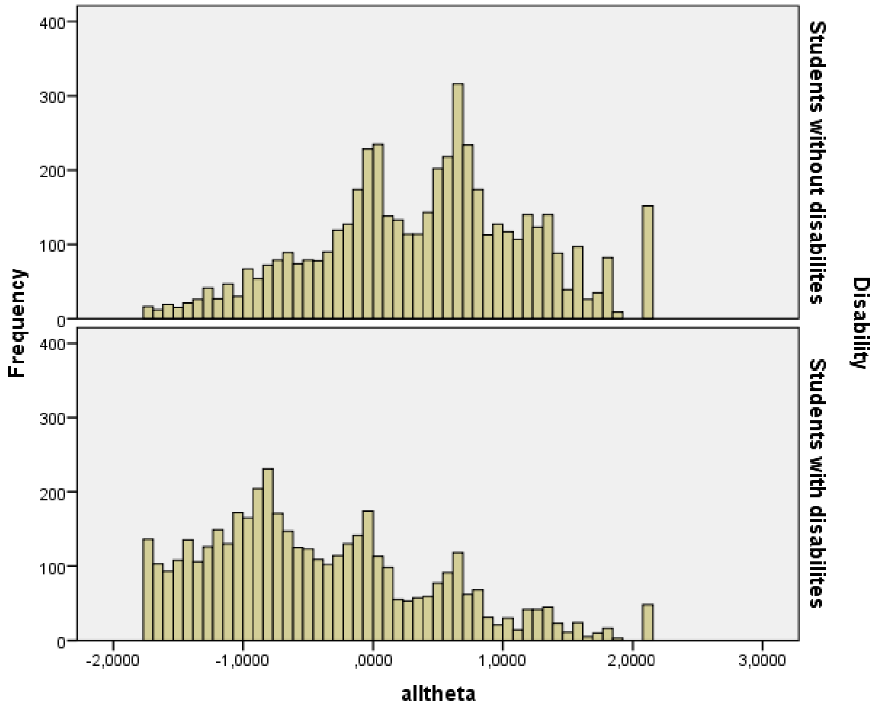

Figure 7. *Estimated θ Levels of Students with and without Disabilities*

Figure 7 shows that the distribution of scores of students with disabilities is skewed to the right, and the scores of those without disabilities have a skewed distribution to the left. This is a visual indication that the average of the achievements of those with disabilities is low and that of those without a disability is high. Table 5 summarizes the results of the independent samples t-test performed to examine whether there is a significant difference between the ability levels of those with and without disabilities.

Table 5. θ means t Test Results According to Disability Status

Group	n	$\bar{X}$	S	df	t	p
Students with disabilities	4410	-,428	,861	9408	-47,479	.000
Students without disabilities	5000	,395	,814			

The results of the independent samples t-test indicated that the θ levels of students with disabilities ($\bar{X}$=-0.428) are significantly lower than the θ levels of students without disabilities ($\bar{X}$ = 0.395). This supports the finding that students with disabilities have lower proficiency than their peers, which underlines the necessity of making arrangements in terms of CAT algorithm stages considering this situation in CAT applications attended by students with disabilities.

How does the optimal CAT algorithm perform at the ability levels of students with or without disabilities who took a large-scale test?

To answer this research question, the optimal CAT conditions found in the first research were studied. For each group, based on the real θ of the whole group (1), the disabled group (2) and the non-disabled group (3), the CAT performances applied under optimal conditions were examined. The results of the CAT simulations carried out with the KL item selection method, ML ability estimation method and *PC* termination rules from the item pool simulated according to the 1000-item 3PL model are presented in Table 6.

Table 6. Performances of the Optimal CAT Condition in Different Groups

Performance indicator	All group	Students without disabilities	Students with disabilities
Simulation time	9410	6.4473	5.581
Number of simulees	12.1018	5000	4410
Mean test length	14.80436	14.776	14.861
Correlation (true θs, estimated θs)	0.9446	0.929	0.936
RMSEA	0.3233	0.3237	0.3242
Bias	0.0038	0.0019	0.0054

As can be seen in Table 6, the averages of test lengths were almost the same in the two distinct groups ($\bar{X}_{sd}$ =14.861; $\bar{X}_{swd}$ =14.776), about 15. Similarly, the correlation between true θs and estimated θs is above 0.92 and is thus quite high. RMSEA values are also the same to the third digit after the decimal point. Although there is no significant difference between the bias values, it can be observed that more biased results are produced in students with disabilities.

DISCUSSION & CONCLUSION

For CAT to be more efficient, item pool designs are frequently discussed in the literature (Belov & Armstrong, 2009; Reckase, 2010; van der Linden et al., 2006). The compatibility of the item pool with the ability levels of the target group is also addressed in such studies. This study examines which CAT procedures are more effective when there is a wide range of ability distributions, by comparing the performances of various CAT procedures. In addition, Turkey Central Secondary Education Exam examined whether the determined procedures would show similar performance in the ability levels of students with disabilities, as a group unfortunately more often of extreme abilities and that CAT will offer advantages in many respects.

Based on the findings of the study, it was observed that the CAT performances of the ML ability estimation method and KL item selection methods were more consistent in different ability ranges in the tests intended to measure students at lower ability levels with precision. In addition, it has been observed that the performances of the PC termination rule and PC termination rule used with a 20-item length limit are similar. In CAT applications where these methods are applied, the average test length is 14.6; the RMSE values are respectively, 0.3172 and 0.3156, and the bias is 0.0112 and -0.0012, respectively. It has been observed that the ML ability estimation method produces more consistent results in different ability ranges. This finding is inconsistent with the findings of Chen, Hou, Fitzpatrick ve Dodd (1997). Chen, et al. (1997) observed that the ML estimation, EAP with a normal prior, and EAP with a uniform prior comparable results methods produced comparable results for a group with normal distribution and a group with skewed distribution. However, it should be kept in mind that the rating scale model was used in this research and the analysis was made with prior distributions. In addition, there was no study examining the performance of CAT procedures in different ability ranges. However, there are studies examining the overall CAT performance of different procedures. Yao (2013) showed the KL item selection method and the PC termination rule among the five optimal procedures in his research examining CAT procedures that are effective in MCAT. This finding supports the research findings.

The methods that were found to perform optimally under the conditions discussed in the study were also tested on real data. First, whether students with disabilities had lower ability levels than those without disabilities, as stated in the literature, was examined (Stone & Davey, 2011). According to the results of the statistical test, the reading comprehension skill discussed in the research is lower in students with disabilities. In the large-scale test designed to answer the main question, it was observed that the performance of the optimal CAT algorithm, which was reached as a result of the research, was high and similar to the ability levels of the students with disabilities and without disabilities. These research findings can be evaluated in the selection of methods in studies that include students with disabilities or students with extreme ability levels. Advantageous aspects of CAT applications include test accommodations that have the potential to provide students with disabilities, no need for extended time accommodations, and more reliable ability estimation (Şenel & Kutlu, 2018a, 2018b); with the use of these methods, more unbiased and valid results can be produced.

In this research, investigations were carried out to show that students with disabilities have lower ability means in the central exam. However, it should not be forgotten that this finding does not mean that all individuals with special needs have lower abilities. It is important to carefully discuss the findings, keeping in mind that the study group is a special group. The results of the research should be interpreted in terms of pointing out that the design of CATs, where individuals with special needs are also tested, should be taken care of.

In this study, θ levels of students with disabilities were obtained from real large-scale test data. However, the CAT item pool was created with a Monte Carlo simulation. In further research, a real item pool and the performance of students with disabilities and other groups in optimal CAT can be compared as a post hoc simulation. There are studies based on such post hoc simulations for the applicability of some large-scale tests as CAT (Seo & Choi, 2018). Similarly, it can be suggested to researchers to compare the power of this optimal CAT application in different ability ranges in a completely real CAT application.

The research focused on measuring the abilities of students with disabilities. However, it is an important result for CAT applications in the field of health, considering that the research has revealed methods that show CAT performance at extreme ability levels similar to that at intermediate ability levels. These CAT procedures may also be preferred in CAT applications for the diagnosis of individuals who fall far below a criterion in terms of certain characteristics such as depression and mental health (Gibbons et al., 2014, 2016).

REFERENCES

Babcock, B., & Weiss, D. J. (2009). Termination criteria in computerized adaptive tests: Variable-length CATs are not biased. *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing*.

Barrada, J. R., Olea, J., Ponsoda, V., & Abad, F. J. (2009). Item selection rules in computerized adaptive testing: Accuracy and security. *Methodology*, *5*(1), 7–17. https://doi.org/10.1027/1614-2241.5.1.7

Belov, D. I., & Armstrong, R. D. (2009). Direct and inverse problems of item pool design for computerized adaptive testing. *Educational and Psychological Measurement*, *69*(4), 533–547. https://doi.org/10.1177/0013164409332224

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

Chen, S. K., Hou, L., Fitzpatrick, S. J., & Dodd, B. G. (1997). The effect of population distribution and method of theta estimation on computerized adaptive testing (cat) using the rating scale model. *Educational and Psychological Measurement*, *57*(3), 422–439. https://doi.org/10.1177/0013164497057003004

Choi, S. W., Grady, M. W., & Dodd, B. G. (2011). A new stopping rule for computerized adaptive testing. *Educational and Psychological Measurement*, *71*(1), 37–53. https://doi.org/10.1177/0013164410387338

Embretson, S., & Reise, S. P. (2000). *Item Response Theory for psychologists*. Lawrence Erlbaum Associates.

Gibbons, R. D., Weiss, D. J., Frank, E., & Kupfer, D. (2016). Computerized adaptive diagnosis and testing of mental health disorders. *Annual Review of Clinical Psychology*, *12*, 83–104. https://doi.org/10.1146/annurev-clinpsy-021815-093634

Gibbons, R. D., Weiss, D. J., Pilkonis, P. A., Frank, E., Moore, T., Kim, J. B., & Kupfer, D. J. (2014). Development of a computerized adaptive test for anxiety. *American Journal of Psychiatry*, *171*(2), 187–194. https://doi.org/10.1176/appi.ajp.2013.13020178

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory, principles and applications*. Springer Science+Business Media. https://doi.org/10.1017/CBO9781107415324.004

He, W., Diao, Q., & Hauser, C. (2013). A comparison of four item-selection methods for severely constrained CATs. *NCME Annual Meeting*, 1–26.

Kezer, F., & Koç, N. (2014). A comparison of computerized adaptive testing strategies. *Eğitim Bilimleri Araştırmaları Dergisi*, *4*(1), 145–174. https://doi.org/10.12973/jesr.2014.41.8

Linacre, J. M. (2000). *Computer-Adaptive Testing: A Methodology whose time has cCome*. Komesa Press.

Lord, F. M. (1980). *Applications of Item Response Theory to practical testing problems*. Routledge.

Magis, D., Yan, D., & von Davier, A. A. (2018). Computerized adaptive and multistage testing with R: Using packages catR and mstR. In *Measurement: Interdisciplinary Research and Perspectives* (Vol. 16, Issue 4). https://doi.org/10.1080/15366367.2018.1520560

Maurelli, V., & Weiss, D. J. (1981). *Factors Influencing the Psychometric Characteristics of an Adaptive Testing Strategy for Test Batteries*.

Ministry of National Education. (2018). *Sınavla öğrenci alacak ortaöğretim kurumlarına ilişkin merkezî sınav başvuru ve uygulama klavuzu [Application guide of central examination for secondary education institutions]*.

Mislevy, R. J., & Bock, R. D. (1982). Biweight estimates of latent ability. *Educational and Psychological Measurement*, *42*(3), 725–737. https://doi.org/10.1177/001316448204200302

Reckase, M. D. (2010). Designing item pools to optimize the functioning of a computerized adaptive test. *Psychological Test and Assessment Modeling*, *52*(2), 127–141. https://psycnet.apa.org/record/2010-17096-001

Riley, B. B., Conrad, K. J., Bezruczko, N., & Dennis, M. L. (2007). Relative precision, efficiency and construct validity of different starting and stopping rules for a computerized adaptive test: The GAIN substance problem scale. *Journal of Applied Measurement*, *8*(1), 48–64. /pmc/articles/PMC5933849/

Sahin, A., & Ozbasi, D. (2017). Effects of content balancing and item selection method on ability estimation in computerized adaptive tests. *Eurasian Journal of Educational Research*, *69*, 21–36. https://doi.org/10.14689/ejer.2017.69.2

Şahin, A., & Weiss, D. J. (2015). Effects of calibration sample size and item bank size on ability estimation in computerized adaptive testing. *Educational Sciences: Theory & Practice*, *15*(6), 1585–1595. https://doi.org/10.12738/estp.2015.6.0102

Segall, D. O. (2004). Computerized adaptive testing. *Encyclopedia of Social Measurement*, 429–438. https://doi.org/10.1016/B0-12-369398-5/00444-8

Seo, D. G., & Choi, J. (2018). Post-hoc simulation study of computerized adaptive testing for the Korean Medical Licensing Examination. *Journal of Educational Evaluation for Health Professions*, *15*, 14. https://doi.org/10.3352/jeehp.2018.15.14

Seo, D. G., & Weiss, D. J. (2015). Best Design for Multidimensional Computerized Adaptive Testing With the Bifactor Model. *Educational and Psychological Measurement*, *75*(6), 954–978. https://doi.org/10.1177/0013164415575147

Şenel, S., & Kutlu, Ö. (2018a). Computerized adaptive testing design for students with visual impairment. *Egitim ve Bilim*, *43*(194), 261–284. https://doi.org/10.15390/EB.2018.7515

Şenel, S., & Kutlu, Ö. (2018b). Comparison of two test methods for VIS: paper-pencil test and CAT. *European Journal of Special Needs Education*, *33*(5), 631–645. https://doi.org/10.1080/08856257.2017.1391014

Şenel, S., & Şenel, H. C. (2018). Bilgisayar tabanlı testlerde evrensel tasarım: Özel gereksinimli öğrenciler için düzenlemeler [Universal design in computer-based testing: Test Accommodations for students with special needs]. In S. Dinçer (Ed.), *Değişen dünyada eğitim* (1st ed., pp. 113–124). Pegem Akademi. https://doi.org/10.14527/9786052412480.08

Stone, E., & Davey, T. (2011). Computer-adaptive testing for students with disabilities: A review of the literature. *ETS Research Report Series*, *2011*(2), i–24. https://doi.org/10.1002/j.2333-8504.2011.tb02268.x

van der Linden, W. J., Ariel, A., & Veldkamp, B. P. (2006). Assembling a computerized adaptive testing item pool as a set of linear tests. *Journal of Educational and Behavioral Statistics*, *31*(1), 81–99. https://doi.org/10.3102/10769986031001081

van der Linden, W. J., & Glas, C. A. W. (2010). *Elements of adaptive testing*. Springer.

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., & Mislevy, R. J. (2000a). *Computerized adaptive testing: A primer*. Routledge.

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., & Mislevy, R. J. (2000b). *Computerized adaptive testing: A primer* (2nd ed.). Lawrence Erlbaum Associates.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*(3), 427–450. https://doi.org/10.1007/BF02294627

Weiss, D. J. (1973). *The stratified adaptive computerized ability test*.

Weiss, D. J. (2011). Better data from better measurements using computerized adaptive testing. *Journal of Methods and Measurement in the Social Sciences*, *2*(1), 1. https://doi.org/10.2458/jmm.v2i1.12351

Yao, L. (2013). Comparing the performance of five multidimensional CAT selection procedures with different stopping rules. *Applied Psychological Measurement*, *37*(1), 3–23. https://doi.org/10.1177/0146621612455687