International Journal of Assessment Tools in Education
2021, Vol. 8, No. 1, 145–155

https://doi.org/10.21449/ijate.735155

Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate **_Research Article_**

The Effect of Item Pools of Different Strengths on the Test Results of Computerized-Adaptive Testing

Fatih Kezer [1,*]

[1]Kocaeli University, Faculty of Education, Department of Educational Sciences, Kocaeli, Turkey

ARTICLE HISTORY

Received: May 10, 2020
Revised: Dec. 17, 2020
Accepted: Jan. 11, 2021

Keywords

Adaptive test,
Item pool,
Item difficulty,
CAT,
IRT.

Abstract: Item response theory provides various important advantages for exams carried out or to be carried out digitally. For computerized adaptive tests to be able to make valid and reliable predictions supported by IRT, good quality item pools should be used. This study examines how adaptive test applications vary in item pools which consist of items with varying difficulty levels. Within the scope of the study, the impact of items was examined where the parameter b differentiates while the parameters a and c are kept in fixed range. To this end, eight different 2000-people item pools were designed in simulation which consist of 500 items with ability scores and varying difficulty levels. As a result of CAT simulations, RMSD, BIAS and test lengths were examined. At the end of the study, it was found that tests run by item pools with parameter b in the range that matches the ability level end up with fewer items and have a more accurate stimation. When parameter b takes value in a narrower range, estimation of ability for extreme ability values that are not consistent with parameter b required more items. It was difficult to make accurate estimations for individuals with high ability levels especially in test applications conducted with an item pool that consists of easy items, and for individuals with low ability levels in test applications conducted with an item pool consisting of difficult items.

1. INTRODUCTION

The measurement and evaluation process plays a critical role in determining whether the qualities targeted to be acquired in education are realized or not. Change has undoubtedly been inevitable in measurement and evaluation just like it has been in every field throughout history. Although paper and pencil tests, which were based on the classical test theory, have been an important part of measurement and evaluation, they have certain important limitations and disadvantages. Item difficulty parameter and item discrimination parameter vary depending on the group from which data were collected; in other words, it varies according to sampling (Lord & Novick, 1968). Another limitation is that individuals' ability levels depend on item parameters. Individuals receive different scores in test batteries with different difficulty levels. One's ability may seem high in an easy test and low in a difficult test. Due to this important limitation, problems may arise in comparing the individual. Even when they could be compared,

CONTACT: Fatih Kezer ✉ fatih.kezer@kocaeli.edu.tr 🖃 Kocaeli University, Faculty of Education, Department of Educational Sciences, Kocaeli, Turkey

ISSN-e: 2148-7456 /© IJATE 2021

because their ability levels are different, their ability scores could cause errors in different sizes (Hambleton, Swaminathan & Rogers, 1991). Tests developed according to traditional approaches and classical test theory usually work better with the individuals with intermediate ability levels (Crocker & Algina, 1986). When few items were designed for individuals with very low- and very high-level abilities, the test ceases to be distinctive for these ability levels, and reliable predictions cannot be made for these extreme ability levels. With existing test designs, it is not possible to know how an individual would perform with a given item set. The limitations of the theory put forth by Spearman in 1905 pioneered the formation of a new theory in 1930s. Item Response Theory (IRT) ties to eliminate limitations due to its strong assumptions (unidimensionality, local independence, model-data fit) and differences in the test algorithm. IRT is also called Latent Trait Theory (Crocker & Algina, 1986). This theory explains with a mathematical function the relationship between an individual's ability level related to the measured characteristic and the answers they give (Embretson & Reise, 2000; Hambleton & Swaminathan, 1989).

The most common item parameters in Item Response Theory are difficulty (*b*), discrimination (*a*), and chance (*c*). Parameter b is the ability (θ) level that corresponds to the point where the individual answers an item correctly with a 50% probability. It is also shown on the same scale as θ (Lord & Novick, 1968). Although it may theoretically take a value between -∞ and +∞, it usually takes in practice a value in the -3 and +3 range. An increase in b denotes that the item is getting more difficult and a decrease indicates that it is getting easier. When parameter b is 0, it denotes a medium-level difficulty. Item discrimination (a) parameter corresponds to the curve on the $\theta=b_i$ point. Theoretically, ranges from -∞ to +∞, however in practice it usually takes a value between 0 and 2. Parameter a can take a negative value, albeit rarely, and this indicates that the item works in the opposite direction. Parameter c denotes the probability of individuals giving a correct answer by guessing.

An important advantage of tem Response Theory is that item and test information functions can be obtained. Item information function shows how much information an item gives of its measured characteristic. Item information is inversely proportional to item error variant (Reid, Kolakowsky-Hayner, Lewis & Armstrong, 2007). A function that takes up a different value in every point of θ is calculated by the equation given below (Baker & Kim, 2004; Hambleton, Swaminathan & Rogers, 1991).

$$I_i(\theta) = \frac{[P_i'(\theta)]^2}{P_i(\theta)Q_i(\theta)}$$

For a three-parameter logistic model, this equation is expressed as follows with item parameters:

$$I_i(\theta) = \frac{2.89\, a_i^2(1-c_i)}{[c_i+e^{1.7\, a_i(\theta-b_i)}][1+e^{-1.7\, a_i(\theta-b_i)}]^2}$$

As parameter a increases and parameter b gets closer to zero, I(θ) value increases as well. Parameter b getting closer to θ is increases I(θ). The total of item information functions gives the test information function that shows how much the test gives information about the measured characteristic (Hambleton, Swaminathan & Rogers, 1991; Reid et al., 2007).

$$I(\theta) = \sum_{i=1}^{n} I_i(\theta)$$

Given the item and test information functions, item characteristics in forming a test is important to be able to have a valid and reliable measuring. IRT provides significant support to measuring processes with its mathematical basis. The invariance characteristic of IRT enables item and test parameters to be independent from the group, and it enables predicted ability levels to be independent from the test. As such, it is possible to compare measuring results of different groups. Being able to calculate the reliability not for a single item but for each of them and for each ability level separately, and also being able to calculate errors separately for each individual enables a shorter test with quality items (Adams, 2005; Crocker & Algina, 1986; Embretson & Reise, 2000; Magnussson, 1966). With its strong mathematical structure, IRT is convenient for various applications. The most important of these are test design, item mapping, test equating, test and item bias studies and computerized adaptive test applications.

In classical tests, a fixed number of items are designed to be applied to all individuals. Adaptive tests, on the other hand, are based on the principle that items appropriate to an individual's ability are used. Thus, the test is cleared of inappropriate items so that it becomes both shorter and more reliable. With the advancement of technology, adaptive tests have begun to be applied more, and computerized adaptive tests (CAT) have gained more importance. In the application of CAT to individuals by selecting items from a large item pool, there are different methods (two-stage testing, self-selecting testing, pyramidal multistage testing, alternating testing, stradaptive testing, multilevel format) (Glas & Linden, 2003; Hambleton & Swaminathan, 1989; Thompson & Weiss, 1980; Vale & Weiss, 1975; Weiss, 1985). Adaptive test strategies are designed to use item information obtained through item information function (Brown & Weiss, 1977; Maurelli & Weiss, 1981; Weiss & Kinsbury, 1984).

The main aim of CAT is to apply the item cluster that gives most information for each individual. To this end, individuals are given different item sets, and based on the answers given to these item sets, an ability estimation is done. Contrary to CTT, CAT is based on IRT and CAT's test logic is based on large item pools item parameters which are known beforehand. Item pool can consist of different item types (Embretson & Reise, 2000; Sukamolson, 2002; Wainer et al., 2000). This testing method requires an item pool which is comprised of items that have high discrimination and that are distributed in a balanced manner on the difficulty-ability level (b-θ) so that it can make estimations for individuals at different ability levels (Geordiadou, Triantafillou & Economides, 2006; Veldkamp & Linden, 2010; Weiss, 1985, 2011). In practice, it is not that easy to form an item pool whose item parameters take value in a large range. In this study, it was examined how estimation of ability changes when item pools consist of items with different characteristics, what kind of differences in the testing would application changing parameter b providing parameters a and c remain in the same range make. Moreover, it examined how estimations of ability changes by creating conditions in which parameter b takes value between narrow and wide ranges and where there is conglomeration at different points from easy to difficult.

2. METHOD

This study is designed as a basic research model in which the psychometric qualities of application results of computerized adaptive tests with items culled from item pools with different difficulty levels, are examined. Basic research refers to those studies that are conducted based on theories, by developing assumptions, testing them, and scientifically interpreting their results (Karasar, 2016).

2.1. Simulation Design

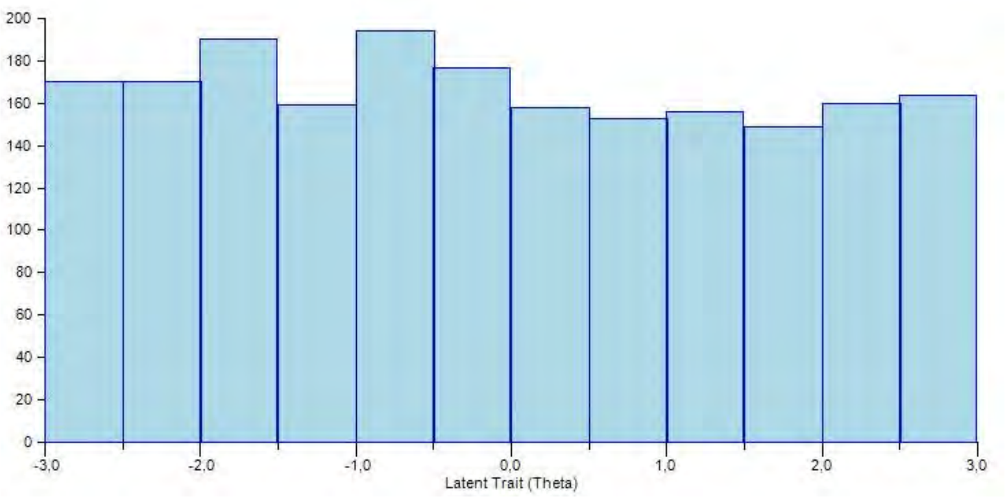
In line with the aim of the study, data were generated in simulation with SIMULCAT Monte-Carlo simulation to compare different item pools. Developed by Kyung T. Han in 2020, SimulCAT is a software to carry out simulated adaptive test applications. When algorithms and codes of practice of adaptive tests are considered, one needs large item pools developed according to item response theory as well as estimated ability parameters from large groups. In this respect, simulative data that could represent each special condition were used in this study. The study was conducted based on a three-parameter model. First, ability parameters were estimated so that they represent a 2000-people group. To estimate the ability, θ (theta) was defined within the -3 and +3 range. Descriptive statistics for estimated ability parameters are presented in Table 1.

Table 1. *Descriptive statistics of ability scores.*

Statistics	Value
N	2000
Mean	-0.073
Median	-0.148
Minimum	-3.000
Maximum	3.000
Range	6.000
Standard Deviation	1.728
Variance	2.985
Skewness	0.091
Std. Error of Skewness	0.055
Kurtosis	-1.188
Std. Error of Kurtosis	0.109

As can be seen in Table 1, mean of the ability parameters generated at (-3, +3) range was found to be -0.073, and its standard deviation 1.718. The same ability parameters (2000-people) were used for all conditions. Distribution related to estimated ability parameters are presented in Figure 1.

Figure 1. *Distribution of ability parameters.*



Eight different conditions were formulated to be able to examine estimated parameters from item pools with different difficulty levels. There are 500 items in each item pool. To see the effect of average difficulty levels, discrimination (a) and chance (c) parameters were defined within the same range so that other conditions remain the same. Parameter a was kept within 0.25 and 2.00, and parameter c within 0.00 and 0.20. Difficulty parameter (b) was defined as a range for each 3 conditions: it was between -3 and +3 for the first condition, -2 and +2 for the second condition and was between -1 and +1 for the third condition. Other than the three ranges, five different conditions were also determined according to average difficulty. In these five different conditions, parameter b was defined as -2.5, -1.5, 0.0, 1.5, and 2.5, respectively, keeping standard deviation as 1.5. Item parameters related to these eight conditions are summarized in Table 2.

Table 2. *Item parameters (defined/generated) for eight different conditons.*

	Defined			Generated					
	b	a	c	b		a		c	
				$\bar{X}$	Sd	$\bar{X}$	Sd	$\bar{X}$	Sd
1st Condition	(-3.0,+3.0)	(0.25,2.0)	(0.0,0.2)	0.017	1.763	1.121	0.498	0.100	0.058
2nd Condition	(-2.0,+2.0)			0.013	1.173	1.139	0.502	0.100	0.058
3rd Condition	(-1.0,+1.0)			0.026	0.594	1.150	0.504	0.101	0.058
4th Condition	$\bar{X}$=2.5 Sd=1.5			2.373	1.567	1.098	0.498	0.098	0.058
5th Condition	$\bar{X}$=1.5 Sd=1.5			1.417	1.536	1.124	0.524	0.101	0.059
6th Condition	$\bar{X}$=0.0 Sd=1.5			0.023	1.406	1.139	0.495	0.102	0.057
7th Condition	$\bar{X}$=-1.5 Sd=1.5			-1.599	1.494	1.138	0.503	0.103	0.057
8th Condition	$\bar{X}$=-2.5 Sd=1.5			-2.577	1.514	1.122	0.522	0.096	0.059

In the adaptive test application design, Maximum Fisher Information (MFI) was used, which is the most common method for item selection management. As initial ability parameter, (-0.5, +0.5) range was determined. Maximum Likelihood Method (MLE) was selected for all conditions as the estimation of ability method. Maximum Likelihood Method is based on selecting the item that gives out most information about an individual. As the termination rule, a common rule was likewise selected for the eight conditions. Standard error which is smaller than 0.30 was determined as the test termination rule. Half the amount of the item pool – 250 items – was decided to be an upper termination rule because too many items would be needed for the estimation of ability if item pool is not appropriate. While conducting the test in inappropriate item pools, the test was stopped when half of the pool is reached. 25 repetitions were made for estimations.

2.2. Data Analysis

In the evaluation of test findings, Root Mean Squared Difference (RMSD) and BIAS values were used. RMSD is a statistic that denotes the difference between estimations of ability (Boyd, Dodd & Fitzpatrick, 2013). BIAS is a difference statistic between the ability parameter average value and its real value. RMSD and BIAS are calculated by using the following formula:

$$RMSD = \sqrt{\frac{\sum_{i=1}^{N}(\hat{\theta}_i - \theta_i)^2}{N}} \qquad BIAS = \frac{\sum_{i=1}^{N}(\hat{\theta}_i - \theta_i)}{N}$$

Moreover, test lengths were also checked in the ability parameter ranges for eight different conditions. The aim was to have a detailed examination of how long the test would take for individuals at different ability levels in the response cluster. Therefore, RMSD and BIAS values at ability ranges were examined.

3. RESULT / FINDINGS

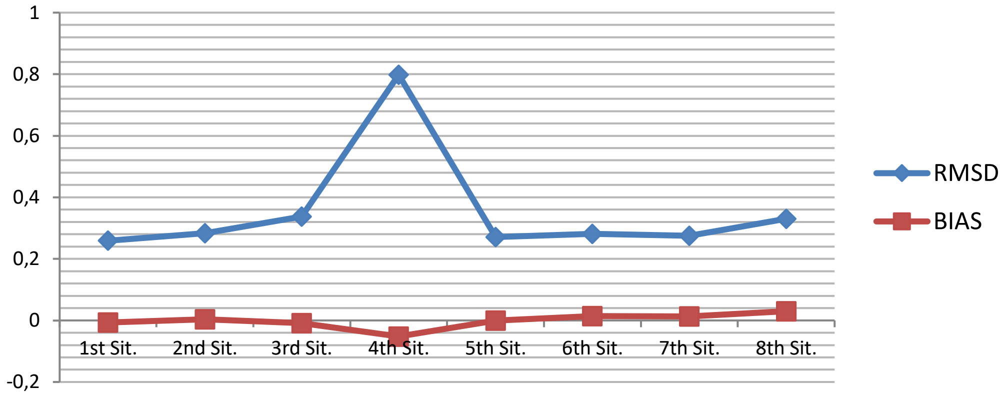
In line with the aim of this study, RMSD and BIAS values for ability parameters obtained from adaptive tests, which were conducted with item pools with different difficulty levels, were calculated and presented in Table 3.

Table 3. *RMSD and BIAS values concerning estimation of ability.*

Condition	b	RMSD	BIAS
1st Condition	(-3.0,+3.0)	0.259	-0.007
2nd Condition	(-2.0,+2.0)	0.283	0.004
3rd Condition	(-1.0,+1.0)	0.338	-0.009
4th Condition	$\bar{X}$=2.5 Sd=1.5	0.798	-0.052
5th Condition	$\bar{X}$=1.5 Sd=1.5	0.271	$-8X10^{-5}$
6th Condition	$\bar{X}$=0.0 Sd=1.5	0.281	0.014
7th Condition	$\bar{X}$=-1.5 Sd=1.5	0.275	0.013
8th Condition	$\bar{X}$=-2.5 Sd=1.5	0.330	0.030

Since 25 repetitions were done in estimations of parameter, obtained results were turned into a report by taking their average. As can be seen in Table 3, RMSD values vary between 0.259 and 0.798. Except for the 4th condition, RMSD values were in a narrower range (0.259-0.338). The lowest RMSD value was obtained, as expected, from the condition in which the difficulty parameters of items in the item pool were between -3 and +3. This value increased when the range of parameter b comparatively narrowed. Apart from when the average was 2.5 in item pools which were formed by considering, the averages of parameter b, no significant difference was detected. Distribution related to RMSD and BIAS values are shown in Figure 2.

Figure 2. *Distribution of RMSD and BIAS values concerning estimations of ability.*



The array of RMSD values according to their size were found to be $RMSD_{Cnd.1}< RMSD_{Cnd.5}< RMSD_{Cnd.7}< RMSD_{Cnd.6}< RMSD_{Cnd.2}< RMSD_{Cnd.8}< RMSD_{Cnd.3}< RMSD_{Cnd.4}$. Similarly, BIAS values concerning different conditions varied absolutely between 0.00008-0.052. Lengths of the simulated adaptive tests were considered separately in θ ranges. Distribution concerning the test lengths are given in Table 4.

Table 4. *Lengths of the simulated adaptive tests.*

θ	N	1st Cnd.	2nd Cnd.	3rd Cnd.	4th Cnd.	5th Cnd.	6th Cnd.	7th Cnd.	8th Cnd.
-3.0<θ<-2.5	170	13.01	55.06	250.00	250.00	147.55	13.57	12.42	11.19
-2.5<θ<-2.0	170	11.91	15.84	182.84	250.00	22.84	12.78	12.56	10.38
-2.0<θ<-1.5	190	11.58	10.91	46.72	134.38	13.39	12.23	12.55	10.58
-1.5<θ<-1.0	159	11.36	11.20	16.42	17.58	13.69	12.08	11.66	11.05
-1.0<θ<-0.5	194	11.23	11.36	11.10	14.29	12.70	12.11	11.45	10.89
-0.5<θ<0.0	177	10.97	11.01	10.89	14.33	12.10	12.58	11.84	11.64
0.0<θ<0.5	158	11.69	10.57	10.65	13.39	11.53	11.73	11.61	14.15
0.5<θ<1.0	153	11.35	10.23	9.88	12.85	11.58	11.30	12.73	17.97
1.0<θ<1.5	156	11.27	11.09	12.97	11.82	11.19	11.21	14.06	27.53
1.5<θ<2.0	149	11.12	10.82	26.77	11.25	11.51	12.67	20.33	138.72
2.0<θ<2.5	160	12.86	15.68	76.79	11.25	11.56	12.46	61.95	250.00
2.5<θ<3.0	164	12.44	35.32	198.61	11.40	11.31	13.12	238.54	250.00

Likewise, RMSD and BIAS values calculated for different conditions for each ability range are given in Table 5.

Table 5. *RMSD and BIAS values according to ability ranges.*

	θ Area	-3.0	-2.5	-2	-1.5	-1	-0.5	0	0.5	1	1.5	2	2.5
	N	170	170	190	159	194	177	158	153	156	149	160	164
1st Cnd.	Bias	-0.02	-0.03	0.00	0.03	-0.03	0.00	-0.01	-0.01	-0.03	-0.02	-0.02	0.06
	RMSD	0.30	0.27	0.25	0.25	0.26	0.27	0.23	0.27	0.26	0.25	0.25	0.26
2nd Cnd.	Bias	-0.08	-0.03	0.00	0.02	-0.01	-0.01	0.00	0.00	-0.01	-0.01	0.09	0.09
	RMSD	0.36	0.31	0.31	0.25	0.24	0.25	0.27	0.23	0.24	0.25	0.35	0.30
3rd Cnd.	Bias	-0.12	-0.05	-0.07	-0.05	-0.03	-0.01	0.01	0.02	0.06	0.08	0.01	0.09
	RMSD	0.60	0.36	0.31	0.28	0.26	0.25	0.26	0.26	0.33	0.32	0.27	0.41
4th Cnd.	Bias	-0.41	-0.08	-0.15	0.00	0.01	-0.02	0.04	0.02	0.01	-0.01	0.02	-0.01
	RMSD	2.03	0.83	1.36	0.30	0.25	0.25	0.25	0.29	0.26	0.25	0.25	0.25
5th Cnd.	Bias	-0.06	0.00	0.02	-0.01	0.01	0.02	0.01	0.01	0.00	0.00	-0.02	0.01
	RMSD	0.40	0.27	0.26	0.26	0.26	0.27	0.25	0.24	0.25	0.26	0.24	0.26
6th Cnd.	Bias	-0.06	0.05	0.02	0.02	-0.03	0.02	0.01	0.05	-0.03	0.01	0.06	0.05
	RMSD	0.46	0.28	0.24	0.25	0.26	0.25	0.25	0.24	0.26	0.24	0.25	0.32
7th Cnd.	Bias	0.00	0.01	0.00	0.02	-0.01	0.02	0.02	0.03	0.01	0.00	0.03	0.03
	RMSD	0.25	0.28	0.26	0.25	0.24	0.24	0.26	0.28	0.27	0.32	0.33	0.32
8th Cnd.	Bias	-0.02	0.02	-0.03	0.03	0.00	0.03	0.05	0.00	0.01	0.05	0.06	0.20
	RMSD	0.26	0.23	0.25	0.25	0.25	0.26	0.29	0.31	0.30	0.32	0.43	0.64

When Table 3 and Table 4 are examined, it can be seen in which ability range item pools with different characters would work more ideally. In the item pool where parameter b is between -3.0 and +3.0 (1st condition), the test was completed, as expected, at a more reasonable time. RMSD and BIAS values were similar and low in each range. In the 6th condition, it was seen that the test length was reasonable for every ability level when parameter b was heaped up around the intermediate difficulty level ($\bar{X}$=0.0, Sd=1.5). Keeping in mind the ability (θ)-difficulty (b) relationship of IRT, it can be said that when difficulty was kept at moderate level, more decisive estimations are done for a large ability range. In the 2nd and 3rd condition in which parameter b was kept within a limited range, it was seen that more items were needed to decisively estimate ability as one moves towards the ends where ability level is high or low. In the 2nd condition, number of items needed at extreme ability levels moved up to 55. In the adaptive test simulation ran in the item pool with parameter b at the (-1.0, +1.0) range, which is a more limited range, (3rd condition), test lengths went outside of acceptable limits in extreme ability levels. The second termination rule of the study – stopping the test when half of the item pool is reached – worked in these three extreme ability levels, and the test was stopped before it could become consistent. This was reflected in RMSD and BIAS values. RMSD value increased to 0.60 in the (-3.0, -2.5) ability range. There was a similar case in item pools which were formed as normal distribution within a certain parameter b. Except for the 6th condition ($\bar{X}$=0.0 Sd=1.5), more items were needed in ranges where parameters b do not correspond to ability levels. As can be seen Table 4, 5th condition-7th condition or 4th condition -8th condition worked adversely and were more decisive in different ability levels. In item pools which were designed by determining parameter b approximately as $\bar{X}$=2.5, the test was stopped by reaching the defined maximum item number without the estimation falling below the standard error value at the $-3<\theta<-2$ range. Similarly, in the 8th condition, the test was stopped as maximum item number was reached at $2<\theta<3$ range. It was observed that RMSD and BIAS values increased in inappropriate ability levels in parallel to test length.

4. DISCUSSION and CONCLUSION

Although classical paper and pencil tests are prevalently used in education and psychology, they give way to electronic exams with the advancements in technology and assessment theories. Item response theory (IRT) provides various important advantages for exams carried out or to be carried out digitally. For computerized adaptive tests to be able to make valid and reliable predictions supported by IRT, good quality item pools should be used (Hambleton, Swaminathan & Rogers, 1991; Weiss, 1985). In adaptive test designs, from 50% to 80% could be saved in test length (Bulut & Kan, 2012; Comert, 2008; Iseri, 2002; Kalender, 2011; Kaptan, 1993; Kezer, 2013; McDonald, 2002; McBride & Martin, 1983; Olsen, Maynes, Slavvson & Ho, 1989; Oztuna, 2008; Scullard, 2007; Smits, Cuijper & Straten, 2011). With CAT, each individual can get a test appropriate for his or her ability level. Moreover, the speed of the test can be adaptive for the individual. Because it is computerized, individuals can take the test at different times where as classical paper and pencil tests everyone should sit in at the same time. Different question formats can be easily used within a test. Test results can be assessed immediately, and test standardization is easier. As an important point, a test that works effectively and properly at every ability level is designed from test that bespeaks to intermediate-level individuals. In order to do a computerized adaptive test that has these advantages, one needs large item pools of which item parameters are estimated beforehand. It is not always easy to write items that has these qualities. Quality of the pool is an important factor that affects efficiency of application. This study examined what kind of results one would get in CAT applications of item pools which have items with different characteristics. Within the scope of the study, the impact of items was examined where the parameter b differentiates while keeping the parameters a and c are kept in fixed range. At the end of the study, it was seen that tests run by item pools with parameter b in the range that matches the ability level end up with fewer items and have a more decisive prediction. Similar studies in literature also underscore when the θ-b relationship is high, more effective CAT applications are carried out (Chang, 2014; Eggen & Verschoor, 2006; Dodd, Koch & Ayala, 1993). When parameter b takes value in a narrower range, estimation of ability for extreme ability values that are not compatible with parameter b required more items. What is more, accurate estimations could not be done with a decent number of items for extreme ability values in much narrower ranges ($-1<b<+1$). Since one could not go below the desired standard deviation, it was difficult to make accurate estimations for individuals with high ability levels especially in test applications conducted with an item pool that consists of easy items, and for individuals with low ability levels in test applications conducted with an item pool consisting of difficult items. These results underline that when generating an item pool in adaptive test applications, one should be incredibly careful.

To make CAT more effective and functional, it can be said that the dimension of the item pool should be as such that would cover all values of b (Chang, 2014). Using items with inappropriate difficulty levels without considering the characteristics of the target group would put adaptive test applications in jeopardy from test length to estimation of ability. The effect of items' levels of difficulty on adaptive test applications can be tested by different item discrimination values at different ability ranges. To this end, examining item parameters would guide teachers and test designers when they form item pools. Moreover, knowing the characteristics of the item pool and its effects could help test designers in constructing correct control mechanisms in test algorithm.

Declaration of Conflicting Interests and Ethics

The author(s) declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author(s).

Authorship contribution statement

Fatih Kezer: All chapters are written by the author.

ORCID

Fatih Kezer https://orcid.org/0000-0001-9640-3004

5. REFERENCES

Adams, R. (2005). Reliabilty as a measurement design effect. *Studies in Educational Evaluation, 31*, 162–172.

Baker, F. B., & Kim, S. H. (2004). *Item response theory: Parameter estimation techniques*. Marcel Bekker Inc.

Boyd, A. M., Dodd, B. & Fitzpatrick, S. (2013). A comparison of exposure control procedures in cat systems based on different measurement models for testlets. *Applied Measurement in Education*, *26*(2), 113-115.

Brown, J. M., & Weiss, D. J. (1977). *An adaptive testing strategy for achievement test batteries* (Research Rep. No. 77-6). University of Minnesota, Department of Psychology, Psychometric Methods Program.

Bulut, O., & Kan, A. (2012) Application of computerized adaptive testing to entrance examination for graduate studies in Turkey. *Egitim Arastirmalari-Eurasian Journal of Educational Research, 49*, 61–80.

Chang, H. H. (2014). Psychometrics behind computerized adaptive testing. *Psychometrika*, 1-20.

Cömert, M. (2008). B*ireye uyarlanmış bilgisayar destekli ölçme ve değerlendirme yazılımı geliştirilmesi [Computer-aided assessment and evaluation analysis adapted to the individual]* [Unpublished master's thesis]. Bahçeşehir University.

Crocker, L., & Algina, J. (1986). *Introduction classical and modern test theory*. Harcourt Brace Javonovich College Publishers.

Dodd, B. G., Koch, W. R., & Ayala, R. J. (1993). Computerized adaptive testing using the partial credit model: effects of item pool characteristics and different stopping rules. *Educational and Psychological Measurement*, *53*(1), 61-77.

Eggen, T. J. H. M., & Verschoor, A. J. (2006). Optimal testing with easy or difficult items in computerized adaptive testing. *Applied Psychologgical Measurement*, *30*(5), 379-393.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum Associates.

Georgiadou, E., Triantafillou, E., & Economides, A. A. (2006). Evaluation parameters for computer adaptive testing. *British Journal of Educational Techonology*, *37*(2), 261–278.

Glas, C. A., & Linden, W. (2003). Computerized adaptive testing with item cloning. *Applied Psychological Measurement*, *27*(4), 247–261.

Hambleton, R. K., & Swaminathan, H. (1989). *Item response teory: Principles and applications*. Kluwer Nijhoff Publishing.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage Publications Inc.

Iseri, A. I. (2002). *Assessment of students' mathematics achievement through computer adaptive testing procedures* [Unpublished doctoral dissertation]. Middle East Technical University.

Kalender, İ. (2011). *Effects of different computerized adaptive testing strategies on recovery of ability* [Unpublished doctoral dissertation]. Middle East Technical University.

Kaptan, F. (1993). *Yetenek kestiriminde adaptive (bireyselleştirilmiş) test uygulaması ile geleneksel kâğıt-kalem testi uygulamasının karşılaştırılması [Comparison of adaptive (individualized) test application and traditional paper-pencil test application in ability estimation]* [Unpublished doctoral dissertation]. Hacettepe University.

Karasar, N. (2016). *Bilimsel araştırma yöntemi [Scientific Research Method]*. Nobel Yayın Dağıtım.

Kezer, F. (2013). *Comparison of the computerized adaptive testing strategies* [Unpublished doctoral dissertation]. Ankara University.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison - Wesley.

Magnusson, D. (1966). *Test theory*. Addison-Wesley Publishing Company.

Maurelli, V. A., & Weiss, D. J. (1981). *Factors influencing the psychometric characteristics of an adaptive testing strategy for lest batteries* (Research Rep. No. 81-4). University of Minnesota, Department of Psychology, Computerized Adaptive Testing Laboratory.

McBride, J. R., & Martin, J. T. (1983). Reliability and validity of adaptive ability tests in a military design. In Weiss, D.J. (Ed.). New horizons in testing: Latent trait test theory and computerized adaptive testing. Academic Press.

McDonald, P. L. (2002). *Computer adaptive test for measuring personality factors using item response theory* [Unpublished doctoral dissertation]. The University Western of Ontario.

Olsen, J. B., Maynes, D. D., Slavvson, D., & Ho, K. (1989). Comparison of paper administered, computer administered and computerized adaptive achievement tests. *Journal of Educational Computing Research*, *5*(31), 311-326.

Öztuna, D. (2008). *An application of computerized adaptive testing in the evaluation of disability in musculoskeletal disorders* [Unpublished doctoral dissertation]. Ankara Üniversitesi Sağlık Bilimleri Enstitüsü.

Reid, C. A., Kolakowsky-Hayner, S. A., Lewis, A. N., & Amstrong, A. J. (2007). Modern psychometric methodology: Applications of item response theory. *Rehabilitation Counselling Bulletin*, *50*(3), 177-178.

Scullard, M.G. (2007). *Application of item response theory based computerized adaptive testing to the strong interest inventory* [Unpublished doctoral dissertation]. University of Minnesota.

Smits, N., Cuijpers, P., & Straten, A. (2011). Applying computerized adaptive testing to the CES-D Scale: A simulation study. *Psychiatry Research*, *188*, 145–155.

Sukamolson, S. (2002). Computerized test/item banking and computerized adaptive testing for teachers and lecturers. http://www.stc.arts.chula.ac.th/ITUA/Papers_for_ITUA_Proceedings/Suphat2.pdf

Thompson, J. G., & Weiss, D. J. (1980). *Criterion-related validity of adaptive testing strategies* (Research Rep. No. 80-3). University of Minnesota, Department of Psychology, Computerized Adaptive Testing Laboratory.

Vale, C. D., & Weiss, D. J. (1975). *A study of computeradministered stradaptive ability testing* (Research Rep. No. 75-4). University of Minnesota, Department of Psychology, Psychometric Methods Program.

Veldkamp, B. P., & Linden. W. J. (2010). Designing item pools for adaptive testing. In Linden, W.J., and Glas, C.A.W.(Eds.). Elements of adaptive testing. Springer.

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., Mislevy, R. J. Steinberg, L., & Thissen, D. (2000). *Computerized adaptive testing: a primer*. Lawrence Erlbaum Associates, Publishers.

Weiss, D. J. (1985). Adaptive testing by computer. *Journal of Consulting and Clinical Psychology*, *53*(6), 774-789.

Weiss, D. J. (2011). Better data from better measurements using computerized adaptive testing. *Journal of Methods and Measurement in the Social Sciences*, *2*(1), 1–27.

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21*(4), 361-375.