Hacettepe Üniversitesi Eğitim Fakültesi Dergisi (H. U. Journal of Education) 34(2): 473-486 [2019]
doi: 10.16986/HUJE.2018042418 [Araştırma Makalesi / Research Article]

Effects of Item Pool Characteristics on Ability Estimate and Item Pool Utilization: A Simulation Study*

Madde Havuzu Özelliklerinin Yetenek Kestirimi ve Madde Havuzu Kullanımına Etkileri: Bir Simülasyon Çalışması

Nagihan BOZTUNÇ ÖZTÜRK**, Melek Gülşah ŞAHİN***

• ***Received:*** *11.12.2017* • ***Accepted:*** *12.07.2018* • ***Published:*** *30.04.2019*

Kaynakça Bilgisi: Boztunç Öztürk, N., & Şahin, M. G. (2019). Madde havuzu özelliklerinin yetenek kestirimi ve madde havuzu kullanımına etkileri: Bir simülasyon çalışması. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 34*(2), 473-486. doi: 10.16986/HUJE.2018042418

Citation Information: Boztunç Öztürk, N., & Şahin, M. G. (2019). Effects of item pool characteristics on ability estimate and item pool utilization: A simulation study. *Hacettepe University Journal of Education, 34*(2), 473-486. doi: 10.16986/HUJE.2018042418

ABSTRACT: Forming an item pool for computerized adaptive testing requires a long and demanding process that may be challenging, both in terms of time and cost. Therefore, one may come across such questions as ‘How should an optimal item pool be?’ and/or ‘How many items should exist in an item pool?’ Although research with regard to the features to exist in an optimal item pool vary, there has been no consensus reached about how big the item pool size should be. In the current study, the effect of different item pool size and item distribution on ability estimation and item pool utilization was analysed. 36 different item pools were generated through SimulCAT software. Using 1,000 simulees, single session CAT environments were simulated and two different termination rules were used in the study. Findings of the study indicated that as the size of the item pool increased to a specific size, the precision of measurement increased and the number of unused items decreased. By examining the results according to *b* parameter, it was found that the effect of *b* parameter distribution over the results decreased.
Keywords: Computerized adaptive testing, item pool quality, item pool size, item pool utilization, ability estimation

ÖZ: Bireyselleştirilmiş bilgisayarlı test uygulaması için madde havuzunun geliştirilmesi uzun ve zahmetli bir süreç gerektirmektedir. Bu süreç hem maddi hem de zaman anlamında yorucu olabilir. Bu nedenle ‘Optimal bir madde havuzu nasıl olmalıdır?, Bir madde havuzunda en az kaç madde yer almalıdır?’ gibi sorularla sıklıkla karşılaşılmaktadır. Optimal bir madde havuzunda bulunması gereken özellikler hakkında yapılan çalışmalar çeşitlilik göstermekle birlikte özellikle madde havuzunun büyüklüğü ile ilgili bir fikir birliği sağlanamamıştır. Bu çalışmada; farklı madde sayısına ve madde dağılımlarına sahip madde havuzlarının yetenek kestirimine ve madde havuzlarının kullanımına etkisi incelenmiştir. Çalışmada 36 farklı madde havuzu SimulCAT yazılımı kullanılarak üretilmiştir. 1000 birey kullanılarak tek oturumluk CAT ortamları simüle ve çalışmada iki farklı sonlandırma kuralı kullanılmıştır. Çalışmanın sonucu genel olarak ele alındığında madde havuzu büyüklüğü belli bir büyüklüğe kadar arttıkça ölçme kesinliğinin arttığı, kullanılmayan madde sayısının azaldığı görülmüştür. Sonuçlara *b* parametresi özelinde bakıldığında madde havuzu büyüdükçe *b* parametresi dağılımının değerler üstündeki etkisinin azaldığı görülmüştür.
Anahtar Sözcükler: Bireyselleştirilmiş bilgisayarlı test, madde havuzu kalitesi, madde havuzu büyüklüğü, madde havuzu kullanımı, yetenek kestirimi

1. INTRODUCTION

Computerized adaptive tests (CAT) are historically based on individualised tests. The individualised tests began with Binet’s intelligence tests in psychometric for the first time

* Some part of this study was presented orally in The 4th International Eurasian Educational Research Congress.

** Inst. Dr., Hacettepe University, Lifelong Learning Center, Ankara-Turkey. e-mail: nagihanboztunc@hacettepe.edu.tr (ORCID: 0000-0002-2777-5311)

*** Inst. Dr., Gazi University, Faculty of Education, Department of Educational Sciences, Division of Assessment and Evaluation in Education, Ankara-Turkey. e-mail: mgulsah@gazi.edu.tr (ORCID: 0000-0001-5139-9777)

(Weiss, 1983) and were followed by the individualisation of group tests by Frederic Lord in the 1970s (Thissen & Mislevy, 2000). Towards the end of the 1980s, as computers became widely used in education, these test formats swiftly started to transform into computer adaptive formats (van der Linden & Glas, 2002). Questions in CAT are asked according to the correct and incorrect answers given by individuals rather than any predetermined order. The primary goal of CAT is to increase the precision of measurement based on a mathematical theory. In addition, it aims to provide a more controlled and a safer testing environment (Hambleton, Swaminathan, & Rogers, 1991).

Reckase (1989) listed four major components of a CAT: the item pool, the item selection procedure, the scoring (ability estimation) procedure, and the stopping rule. Item exposure control and content balancing have recently been widely studied to constrain the item selection in order to select items, not only by their statistical characteristics, but also by content specifications and security concerns (Gu & Reckase, 2007). The point of CAT is as a test algorithm in which key components take place in a certain order. To illustrate; an examinee sits at a computer to take a test. The computer is preloaded with the item pool (which includes psychometric data on each item), and a specific starting point will have been determined for the examinee. An item is selected for this starting point. After the examinee answers the first item, it will be scored and will give an estimate of examinee's ability (θ). The termination criterion will then be evaluated; unless satisfied, another item will be selected for the examinee to answer, then the examinee's score (θ) is updated, and the termination criterion will be evaluated once again (Thompson & Weiss, 2011). This condition repeats until the condition of the selected termination criterion is satisfied.

Each component of CAT is respectively important. Even the changing of only one component affects the amount of error in the ability estimation of the individuals. For computerized adaptive tests to work well, they must have an item pool with sufficient numbers of good quality items (He & Reckase, 2013). Characteristics of the item pool, which includes size, item parameter distributions, and content coverage, directly affect CAT measurement efficiency and test score validity (Segall, 2004). Many researchers have stated that, in developing item pools for CATs, the item pool size, the distribution of item parameters and practical considerations such as content distribution and item exposure issues are extremely important (He & Reckase, 2013).

Forming an item pool is a quite demanding process. Writing an item, checking and pretesting it requires considerable time and is a costly process (Veldkamp & van der Linden, 2010; Zhou & Reckase, 2014). For this reason, it is important to determine the features of the item pool right from the beginning.

However, in spite of CAT research having started to exist during the late 1960s, it was not until the early 2000s that the field started to see the emergence of relevant research solely dedicated to item pool design for CATs (He & Reckase, 2013). In the 1990s and 2000s, CAT research literature expressed that item pool quality in realising CAT's measurement quality is important (Flaugher, 2000; Gorin, Dodd, Fitzpatrick, & Shieh, 2005; Wang & Kolen, 2001; Wang & Vispoel, 1998; Xing & Hambleton, 2004).

In the literature, the features that an optimal item pool should have are stated similarly. An optimal item pool should be determined by the other components of the CAT, namely test length, expected distribution of the examinee population, ability estimation and item selection procedures, and target item exposure and overlap rates (Bergstrom & Lunz, 1999). An optimal item pool is defined as one that can always provide optimal items that satisfy the expected specifications of a CAT program in its implementation process. To achieve this, an item pool must have a sufficient number of items and a distribution matching the target population (Boyd, Dodd, & Choi, 2010). He and Reckase (2013), in order for the optimal item features to reflect the intended test and examinees characteristics, all item pool designs need to be to the design of the testing programs, target examinee population, and a specific test purpose. Parshall, Spray, Kalohn, and Davey (2002) stated that an item pool should include a sufficient number of high quality items that are targeted to the examinee population. Ariel, van der Linden, and Veldkamp (2006) stated that an optimal item pool should consist of a maximal number of combinations of items that (a) meet all content specifications for the test, and (b) are most informative at a series of ability levels reflecting the shape of the distribution of the ability estimates for a population of examinees. Wise (1997) stated that a high-quality item pool should contain a sufficient number of useful items that allow for efficient, informative testing at important levels of proficiency. The size and item difficulty distribution of an item pool depends on the CAT specifications and the examinee population (Reckase, 2010). To summarise, it is expected that there should exist a sufficient number of items in an optimal item pool and that it should have the desired psychometric features relevant to the purpose of the test and target population.

How many items should be in a pool is another question often asked during item pool design. Ideally, the more items the better, because it allows more choices in test assembly, and rarely do the same items exist in tests repeatedly. Within larger item pools, it is difficult for examinees to memorise answers. This may cause a problem in situations where learners have access to the item pool. Millman and Arter (1984) state that larger item pools also mean more that items that match content, item format, and statistical requirements are available. Gu and Reckase (2007) stated the caveats as being: (1) the items added to the pool should be well written, content valid, and statistically fit; and (2) the total number of items should be manageable and easily retrievable.

According to the related previous researches on the CAT, the features of an optimal item pool should include a sufficient number of high quality items that are targeted to the examinee population and CAT specifications (Ariel et al., 2006; Bergstrom & Lunz, 1999; Boyd et al., 2010; He & Reckase, 2013; Parshall et al., 2002; Reckase, 2010; Wise,1997) Besides, it could be concluded according to the related literature that there have been different findings on the item pool size (Chen, Ankenmann, & Spray, 2003; Flaugher, 2000; Stocking, 1994; Urry, 1977). The fact that the item pool is an important component in CAT and that the item pool characteristics and sizes used in the research are different, led the researchers to work on this issue. The current study aims to examine the item pools which have psychometric features and item pool sizes suggested in the related literature. In the study, how the item pool size and psychometric features impact different item termination, ability estimation in the tests by different numbers of contents and pool utilization was analysed.

2. METHOD

The current study aims to examine how item pools with different psychometric features and sizes impact on ability estimation in a CAT that uses 3PLM, and so as to identify which item pools are utilized more efficiently. To find out about the sub-problems determined in accordance with this purpose, the situations determined within the study will be simulated. Therefore, simulation model was used for the current study.

Using a CAT facilitates the use of simulations to aid in the design and maintenance of the assessment. For example, simulations can be set up to test how the assessment performs when different criteria are used to determine, the start rule, stop rule, item selection, and other CAT characteristics (Jacobsen et al., 2011).

2.1. Item Pools and Simulees

For the current study, the SimulCAT simulation software tool (Han, 2011) was used both to generate data and for CAT simulation. Six different item pool sizes, two different *b* parameter distributions and three different contents were used. In conclusion, 36 (6x2x3) different item pools in total were included in the study.

Sizes: The sizes of the item pools were determined as 100, 140, 240, 300, 500, and 960 according to the literature (Chen et al., 2003; Flaugher, 2000; Stocking, 1994; Urry, 1977; Weiss, 1985).

Characteristics: The *a* parameters were determined in the [0.50, 2.00] interval, and the c parameters were determined in the [0.05, 0.20] interval with uniform distribution in the all item pools. The *b* parameters of these item pools had the same distributions; however, the *b* parameter distributions differed. The item pools defined as normal in the study had a uniform *b* parameter distribution ($N(1,1)$), whereas in the uniform item pools the *b* parameters were determined in the [-3, 3] interval with uniform distribution.

Content Areas: The number of content areas are 1, 2, and 4. In the simulations that have a contents of 2, the content distribution is 50%, and in other simulations with a contents of 4 the content distribution is determined as 25%. While generating the item pools, item features which belong to each content areas were designed in accordance with the item distributions stated in characteristics section.

Simulees: The ability distributions for the simulees in the sample group of the CAT were derived from the normal distribution ($N(0,1)$), with a mean of 0 and a standard deviation of 1. The size of the sample was determined as 1,000.

2.2. Components of CAT

The CAT administration based on dichotomously scored items was simulated in this research. Therefore, the 3PLM was employed. For the selection of the first item, the simulees' initial theta estimate was set to zero. To estimate ability, the estimation method of Expected A Posteriori (EAP) was used. Fixed test length and variable length were chosen as a termination rule. The fixed test length was set to 20 and the variable length was set to change when interim theta estimates became smaller than 0.02. The simulation was designed to simulate 1,000 simulees attending a single CAT session during a single test time slot. The Maximum Fisher Information (MFI) method was used for item selection. The Fade-Away method (Han, 2012), with a target exposure rate of 0.20, was chosen as the item exposure control method. The weight method (Kingsbury & Zara, 1989) was used for balancing the content areas, and 25 replications were used in the research. The mean was calculated for the results obtained and the analyses performed.

2.3. Data Analysis

After the CAT simulation, the fidelity coefficient, RMSE, bias, average absolute difference, mean of standard error of estimation, and the number of unused items were calculated for each condition. Pearson's Product Moments Correlation was employed in calculating the fidelity coefficient. The formulas of RMSE, Bias, and the average absolute difference (AAD) were as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{\theta}_i - \theta_i)^2}{n}}$$

$$Bias = \frac{\sum_{i=i}^{n}(\hat{\theta}_i - \theta_i)}{n}$$

$$AAD = \frac{\sum_{i=1}^{n}\left|\hat{\theta}_i - \theta_i\right|}{n}$$

$\hat{\theta}_i$ represents the estimated level of ability for person i, θ_i represents the known level of ability for person *i*, and *n* represents sample size in these formulas (Boyd, 2003; Davis, 2002).

Standard Error of Estimation: SEE for the final theta estimates were used for the graphs. After 25 replications, SEE values obtained from SimulCAT software were averaged on the basis of theta. The change of the average SEEs was shown in diagrams according to theta levels. Theta levels were discretised while creating the diagrams. 12 intervals were obtained from 0.50 intervals.

Number of Unused Items: In order to determine the number of unused items, the use of items from 25 replications were analysed. Items unused in any replication were named as ‘unused items’. The number of unused items was divided into 25, and their percentage and average number were calculated. Theta levels were discretised while creating the diagrams. 12 intervals were obtained from 0.50 intervals.

Test Length: The average test length for each individual over 25 replications was calculated for each condition. Theta levels were discretised while creating the diagrams. 12 intervals were obtained from 0.50 intervals.

3. RESULTS

The findings of the current study are given under subsections entitled Measurement Precision, Standard Error of Estimation, Test Length, and Number of Unused Items.

3.1. Measurement Precision

To evaluate the measurement precision, the average values of fidelity, RMSE, bias and average absolute difference indicators were calculated. The findings, based upon the termination rule, fixed-length test (*n* = 20) and variable-length test (standard error of ability estimation threshold [*SE* = 0.020] for the last three items) are shown in Appendix 1 and Appendix 2.

3.1.1. *Fixed-Length Test*

The fidelity, RMSE, bias and AAD values were calculated according to the fixed-length termination rule (see Appendix 1). Accordingly, in the item pools with a content of 1, where the *b* parameter had uniform distribution, the highest fidelity value, lowest RMSE, bias and AAD values were in the item pool with 960 items. However, the lowest fidelity value, and the highest RMSE and AAD values were in the item pool with 100 items. Similar values are seen as valid in the conditions with content areas 2 or 4. When the number of items increased, the precision of the measurement became higher. This condition did not vary as the number of content increased. However, the number of content varied from 1 to 4, and fidelity value usually decreased, whereas RMSE, bias and AAD values increased.

In the item pools where the *b* parameter had normal distribution, the highest fidelity value for a content area 1, and the lowest RMSE and AAD values were found in the pool with 960 items. However, the lowest fidelity value, the highest RMSE and AAS value were found in the pool with 100 items. Similar values are valid in the conditions with content areas 2 or 4. When the number of items increased, the precision of the measurement became higher. This condition did not vary as the number of content increased. However, the number of content varied from 1 to 4, and fidelity value usually decreased, whereas RMSE, bias, and AAD values increased.

3.1.2. *Variable Length Test*

The fidelity, RMSE, bias, and AAD values were calculated where the variable-length termination rule was chosen (see Appendix 2). Accordingly, in the item pools with a content area 1, where the *b* parameter had uniform distribution, the highest fidelity value, the lowest RMSE, bias, and AAD values were in the item pool with 960 items. However, the lowest fidelity value, the highest RMSE and AAD values were in the item pool with 100 items. Similar values are valid in the same conditions where the content areas were 2 or 4. When the number of items increased, the precision of the measurement became higher. This condition did not vary as the number of content increased. However, the number of content area varied from 1 to 4, and fidelity value usually decreased, whereas RMSE, bias, and AAD values increased.

In the item pools where the *b* parameter had normal distribution, the highest fidelity value in content 1, the lowest RMSE and AAD values were found in the pool with 960 items. However, the lowest fidelity value, the highest RMSE and AAS value were found in the pool with 100 items. Similar values are valid in conditions with content areas 2 or 4. When the number of items increased, the precision of the measurement became higher. This condition did not vary as the number of content increased. However, the number of content varied from 1 to 4, and fidelity value usually decreased, whereas RMSE, bias, and AAD values increased.

When the item pools with different *b* parameter distributions were analysed (see Appendix 1 and Appendix 2), the difference between the fidelity, RMSE, bias, and AAD values, which belong to the item pools with a limited number of items differ more than the difference between the fidelity, RMSE, bias, and AAD values which belong to item pools with a large number of items. In other words, when the item pool became larger, the effect of *b* parameter distribution on values decreased.

Bias, which changed most but did not show any linear change among the measurement precision, had a tendency to up to down when the number of items increased. It was found that measurement precision increased when the variable length test was adopted.

3.2. Standard Error of Estimation for Final Theta Estimation

In Figure 1, the change of standard error of estimation can be seen according to theta by different termination rule, *b* parameter distribution and different size of item pools.

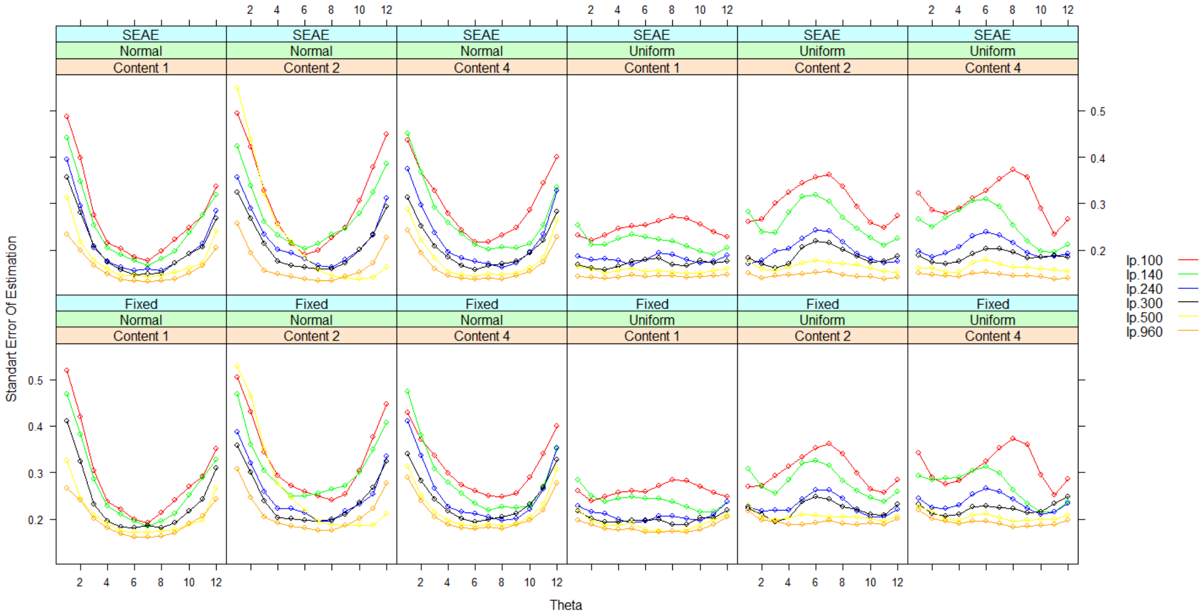


Figure 1. Standard error of estimation distributions for each condition in fixed-length and variable-length termination rules

It can be concluded from Figure 1 that the standard error is quite high and the *b* parameter had normal distribution; whereas it had the lowest value in the intervals in which theta is around 1. Especially, the error values between [-3, -2] were higher than the error values within [2, 3].

When the number of items increased, it was found that SEE values decreased dramatically, and the SEE variance between the high and low extreme values decreased. When different test termination rules were compared, especially in the item pool with 100 items, it was found that there was no significant difference. However, when the number of items in the item pool increased, so did the SEE values in the larger item pools.

In the item pools where the *b* parameter had uniform distribution, standard error within small item pools, especially in pools with 100 and 140 items, and different theta levels varied considerably. However, in the larger item pools; especially in pools with 500 or 960 items, with different theta levels, the standard errors were seen to be similar.

3.3. Test Length

Figure 2 shows the change of test lengths according to theta level in different item pool sizes and different distributions of *b* parameter where the variable-length test termination rule was implemented.

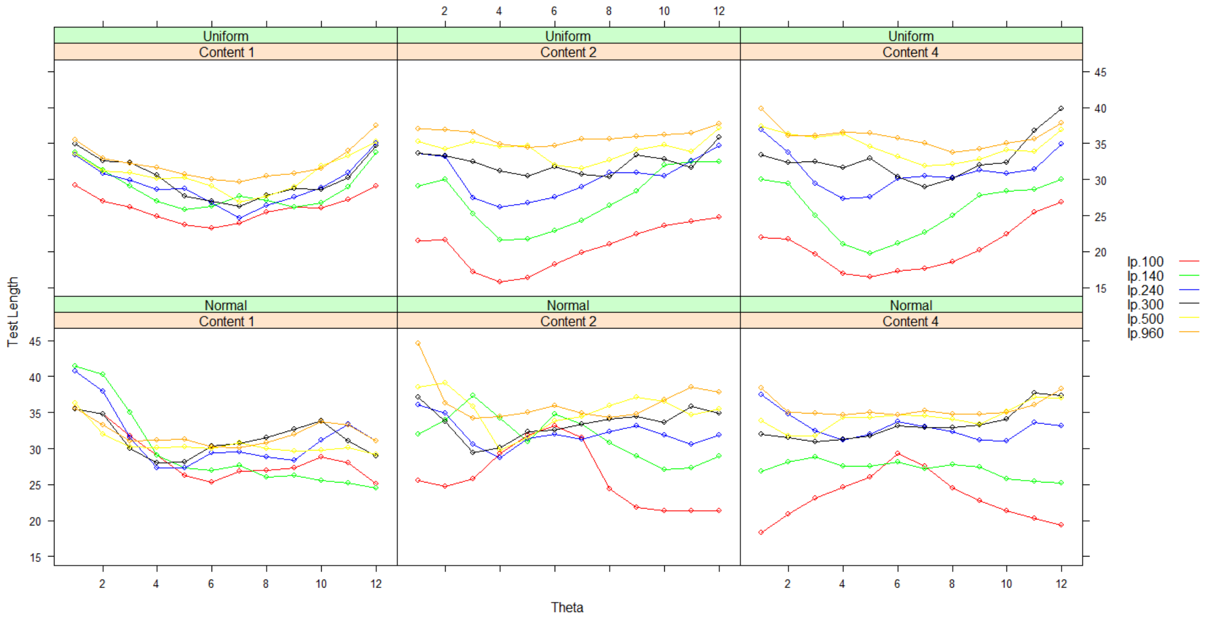


Figure 2: Average test lengths obtained for each condition in variable length termination rule

When Figure 2 is analysed, it can be concluded that in item pools where the *b* parameter had normal distribution and the test had a content area 1, the items used within low theta levels were between 35-45. In addition, the item pools usually displayed a similar manner. However, when the number of content increased, test lengths varied according to the size of item pools. It was found that individuals showing a moderate level with an item pool size with 100 items used more items for theta prediction. When the size of the item pool became larger, so did the test lengths.

In the item pools where the *b* parameter had normal distribution, when the test had a content area 1, test lengths were found to be similar. It was seen that the shortest tests had an item pool size with 100 items, whereas the longest tests were implemented within the largest item pools. When a content area 2 or 4 was analysed, it was seen that the difference among the item pools increased. In pools sizes with 100, 140, and 240 items at a certain theta level, short tests were implemented.

It was found that tests of different lengths were used according to theta levels when the item pools were evaluated based on *b* parameter distribution. However, in the larger item pools, similar test lengths were implemented.

3.4. Number of Unused Items

3.4.1. *Fixed Length*

Table 1 presents the number and percentages of unused items according to item pools of different sizes, when the fixed-length termination rule was implemented, and *b* parameter had a uniform and normal distribution.

Table 1. Number & percentage of unused items for each condition in fixed-length termination rule

***b* Parameter Distribution**	**Item Pool Size**	**Number of Test Content**					
		1		**2**		**4**	
		F	**%**	**f**	**%**	**f**	**%**
Uniform	100	16	16.00	0	0.00	0	0.00
	140	38	27.14	2	1.43	3	2.14
	240	108	45.00	35	14.58	32	13.33
	300	149	49.67	72	24.00	71	23.67
	500	322	64.40	225	45.00	213	42.60
	960	738	76.88	577	60.10	574	59.79
Normal	100	19	19.00	0	0.00	0	0.00
	140	45	32.14	0	0.00	0	0.00
	240	120	50.00	37	15.42	32	13.33
	300	166	55.33	67	22.33	68	22.67
	500	342	68.40	229	45.80	194	38.80
	960	763	79.48	562	58.54	547	56.98

Table 1 shows that in item pools where the *b* parameters had both uniform and normal distributions, and for a content area 1, it was found that unused items existed in all the item pools. When the *b* parameter had a uniform distribution and the number of content area was 2 or 4, all the items were used in the size with 100 item pool. However, when the *b* parameter had normal distribution, all items were used in the pools of 100 and 140 items.

When the size of the item pool became larger, so did the number of unused items. A single content item pool with 960 items, where the *b* parameter had normal distribution and which was comprised of 80% unused items, included more unused items than other item pools. It was found that these values decreased when the number of content increased. In item pools with similar conditions, 57% of the items remained unused.

3.4.2. *Variable Length*

Table 2 presents the number and percentages of unused items in item pools with different sizes, when the variable-length termination rule was implemented and *b* parameter had uniform and normal distribution. These items were not used in any replications.

Table 2: Number & percentage of unused items for each condition in variable-length termination rule

b Parameter Distribution	Item Pool Size	Number of Test Content					
		1		2		4	
		F	%	f	%	f	%
Uniform	100	0	0.00	0	0.00	0	0.00
	140	3	2.14	0	0.00	0	0.00
	240	42	17.50	0	0.00	0	0.00
	300	79	26.33	2	0.67	6	2.00
	500	228	45.60	91	18.20	100	20.00
	960	579	60.31	412	42.92	396	41.25
Normal	100	0	0.00	0	0.00	0	0.00
	140	3	2.14	0	0.00	0	0.00
	240	51	21.25	0	0.00	0	0.00
	300	85	28.33	5	1.67	7	2.33
	500	243	48.60	104	20.80	95	19.00
	960	628	65.42	388	40.42	393	40.94

Table 2 shows that in the item pools where the *b* parameters had both uniform and normal distributions and when they had a content area 1, it was found that unused items existed in all item pools except for the item pool with 100 items. When the number of content size was 2 or 4, all the items were used in the pools with 100, 140, and 240 items. However, when the size of the item pool was 300, almost no items were used.

When the size of item pool became larger, so did the number of unused items. A single content item pool with 960 items, where the *b* parameter had normal distribution and which was comprised of 66% unused items, included more unused items than other item pools. It was found that these values decreased when the number of content increased. In item pools with similar conditions, nearly 40% of the items remained unused.

4. DISCUSSION and CONCLUSION

In this study, the effect of different item pool size and item distribution on ability estimation and item pool utilization was analysed. A summary of the results are as follows:

- When the effect of item pool size on measurement precision was analysed, it was concluded that the measurement precision, both for fixed and variable length termination rules, increases when the item pool size becomes larger. Generally, the lowest RMSE, bias, and AAD values were obtained from the largest item pools; whereas the highest values were obtained from the smallest item pools.

  When the item pools with different *b* parameter distributions were analysed in similar conditions, the difference between the fidelity, RMSE, bias, and AAD values, which belonged to item pools with limited numbers of items were higher than the difference between the fidelity, RMSE, bias, and AAD values of item pools with a large number of items. In other words, when the item pool becomes larger, the effect of *b* parameter distribution on values decreases.

  Bias, which changes most but does not show any linear change among the measurement precision, had a tendency to wax and wane when the number of items increased.

However, the measurement precision increased when the termination rule was standard error ability estimation threshold.

- When the ability level distributions of standard error of estimation were analysed, SEE levels were high in the extreme values and the lowest around $\theta = 1$ in the item pools where the b parameter had normal distribution. As the b parameters had a $N(1,1)$ type distribution, the SEE level distributions assumed those values.

  When the size of the item pool became larger, the difference between SEE values decreased. In the item pools where the b parameter had uniform distribution, the distribution of SEE values varied in smaller item pools; whereas they were similar in larger item pools.

- When the number of contents increased, so did the SEE values both in uniform and normal distributions, and this condition impacted more on the item pools with 100 and 140 items. The distribution of SEE was similar in item pools where the b parameter had normal distribution, although the number of contents increased. However, in the item pools with 100, 140, 240, and 300 items, where the b parameter had uniform distribution, SEE values were high for the medium ability levels.

  When the SEE distributions in different termination rules were analysed, although the number of contents increased in the item pools where the b parameter had normal distribution, SEE values did not change much. However, in item pools where the b parameter had uniform distribution, and when the number of contents increased, they changed in the smaller item pools. This condition, where the b parameter had uniform distribution, existed when the item pools were larger than 500 items. It can be concluded that if the number of contents are more than 1 and the size of the item pool is larger than 500, then the distribution levels will not vary considerably. When the number of content was 1, this condition existed in item pools with at least 300 items.

- According to the variable length termination rule and the theta levels of the tests implemented to the examinee, in item pools where the b parameter had normal distribution and a content of 1, the distribution of test lengths were found to be similar to the lengths in the item pools with different sizes. This was similar for item pools where the b parameter had uniform distribution. However, when the number of contents increased, so did the test lengths; especially in the smaller item pools. In the item pools where the b parameter had normal distribution, longer tests were implemented in the smaller item pools when the theta level was medium. In addition, the average test length increased as the item pool size became larger.

- When the number of unused items were analysed in CAT implementation, regardless of the distribution of b parameter in which the fixed-length termination rule was applied, it can be concluded that unused items existed when the number of content was 1. However, when the number of content increased, the number of unused items decreased. The reason for this change is that item exposure methods were applied while balancing the content; therefore, the number of unused items decreased. When the variable length termination rule was applied, an increase in the number of content resulted in the decrease in the number of unused items. When the size of the item pool became larger, the number of unused items also increased. When the results from the two different termination rules were compared, it could be concluded that the number of unused items which are not used in fixed-length tests is higher.

In the current study, the effect of the b parameter on two different types of distribution of b parameter (normal and uniform) and on the usage of the item pool was examined. In future studies, the b parameter may be skewed to the right and/or to the left, or it may be of normal distribution in order to produce a different mean and standard deviation. The scope of the current study addresses only three different sizes of content; whereas a different research design could be created by simulating the larger test content. Simulation data was used in the current study. By studying all of these variables, studies on real data could be achieved.

The current study was limited to the methods specified in CAT components. Future studies could analyse using different starting, termination, item selection, content balancing, item exposure control, and ability estimation methods.

5. REFERENCES

Ariel, A., van der Linden, W. J., & Veldkamp, B. P. (2006). A strategy for optimizing item-pool management. *Journal of Educational Measurement, 43*(2), 85-96.

Bergstrom, B. A., & Lunz, M. E. (1999). CAT for certification and licensure. In F. Drasgow, & J. B. Olson-Buchanan (Eds.), *Innovations in computerized assessment* (pp. 67-91). Mahwah, NJ: Lawrence Erlbaum.

Boyd, A., Dodd, B., & Choi, S. (2010). Polytomous models in computerized adaptive testing. In M. L. Nering, & R. Ostini (Eds.), *Handbook of polytomous item response theory models* (pp. 229-255). New York: Routledge.

Boyd, M. A. (2003). *Strategies for controlling testlet exposure rates in computerized adaptive testing systems* (Unpublished Doctoral dissertation). The University of Texas, Austin.

Chen, S.Y., Ankenmann, R.D., & Spray, J.A. (2003). The relationship between item exposure and test overlap in computerized adaptive testing. *Journal of Educational Measurement, 40*(2), 129-145.

Davis, L. L. (2002). *Strategies for controlling item exposure in computerized adaptive testing with polytomously scored items* (Unpublished Doctoral dissertation). The University of Texas at Austin, Austin.

Flaugher, R. (2000). Item pools. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (pp. 37-59). Mahwah, NJ: Lawrence Erlbaum.

Gorin, J. S., Dodd, B. G., Fitzpatrick, S. J., & Shieh, Y. (2005). Computerized adaptive testing with the partial credit model: Estimation procedure, population distributions, and item pool characteristics. *Applied Psychological Measurement*, *29*(6), 433-456.

Gu, L., & Reckase, M. D. (2007). Designing optimal item pools for computerized adaptive tests with sympson-Hetter exposure control. *Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing.*

Hambleton, R.K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. New York: Sage publication.

Han, K. T. (2011). *User's manual: SimulCAT.* Retrieved 01 November 2017 from http://www.umass.edu/remp/software/simcata/simulcat/SimulCAT_Manual.pdf.

Han, K. T. (2012). An efficiency balanced information criterion for item selection in computerized adaptive testing. *Journal of Educational Measurement, 49*(3), 225-246.

He, W., & Reckase, M. D. (2013). Item pool design for an operational variable length computerized adaptive test. *Educational and Psychological Measurement, 74*(3), 473-494.

Jacobsen, J., Ackermann, R., Eguez, J., Ganguli, D., Rickard, P., & Taylor, L. (2011). Design of a computer-adaptive test to measure English literacy and numeracy in the Singapore workforce: considerations, benefits, and implications. *Journal of Applied Testing Technology, 12*(SI), 1-26.

Kingsbury, G. G., & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education, 2*(4), 359-375.

Millman, J., & Arter, J. A. (1984). Issues in item banking. *Journal of Educational Measurement, 21*(4), 315-330.

Parshall, C., Spray, J., Kalohn, J., & Davey, T. (2002). *Practical considerations in computer-based testing.* New York: Springer Verlag.

Reckase, M. D. (1989). Adaptive testing: The evolution of a good idea. *Educational measurement: Issues and practice, 8*(3), 11-15.

Reckase, M. D. (2010). Designing item pools to optimize functioning of a computerized adaptive test. *Psychological Test and Assessment Modeling, 52*(2), 127-141.

Segall, D. O. (2004). *Computerized Adaptive Testing*. Encyclopaedia of Social Measurement, Academic Press. Retrieved from http://iacat.org/sites/default/files/biblio/se04-01.pdf.

Stocking, M. L. (1994). Three practical issues for modern adaptive testing item pools (ETS Research Report No. 93-2). Educational Testing Service: Princeton, NJ.

Thissen, D., & Mislevy, R. J. (2000). Testing algorithms. In H. Wainer (Ed.), *Computerized Adaptive Testing: A Primer* (2nd ed., pp. 101-135). London: Routledge.

Thompson, N. A., & Weiss, D. J. (2011). A Framework for the development of computerized adaptive tests. *Practical Assessment, Research & Evaluation, 16*(1).

Urry, V. W. (1977). Tailored testing: a successful application of latent trait theory. *Journal of Educational Measurement, 14*(2), 181-196.

van der Linden, W. J., & Glas, C. A. V. (2002). *Computerized adaptive testing: theory and practice.* USA: Kluwer Academic.

Veldkamp, B. P., & van der Linden, W. P. (2010). Designing item pools for adaptive testing. In W. J. van der Linden, & C. A. Glas (Eds.), *Elements of Adaptive Testing* (pp. 231-245). New York: Springer.

Wang, T., & Kolen, M. J. (2001). Evaluating comparability in computerized adaptive testing: Issues, criteria and example. *Journal of Educational Measurement, 38*(1), 19-49.

Wang, T., & Vispoel, W. P. (1998). Properties of ability estimation methods in computerized adaptive testing. *Journal of Educational Measurement, 35*(2), 109-135.

Weiss, D. J. (1983). *New Horizons in Testing*. New York: Academic Press.

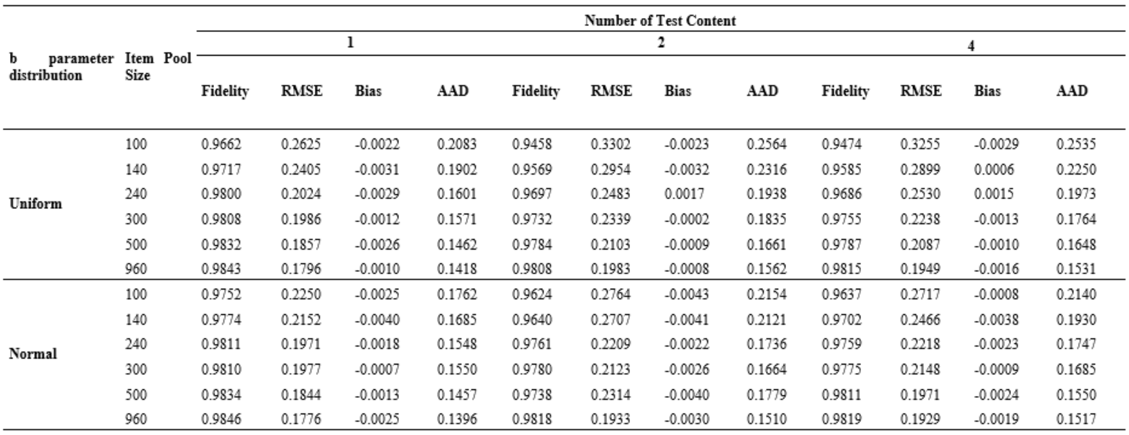
Weiss, D. J. (1985). Adaptive testing by computer. *Journal of Consulting and Clinical Psychology, 53*(6), 774-789.

Wise, S. L. (1997). *An Evaluation of the Item Pools Used for Computerized Adaptive Test Versions of The Maryland Functional Tests.* A Report Prepared for the Assessment Branch of the Maryland State Department of Education. Retrieved 10 March 2017 from https://marces.org/mdarch/pdf/M032045.pdf.

Xing, D., & Hambleton, R. K. (2004). Impacts of test design, item quality, and item bank size on the psychometric properties of computer-based credentialing examinations. *Educational and Psychological Measurement, 64*(1), 5-21.

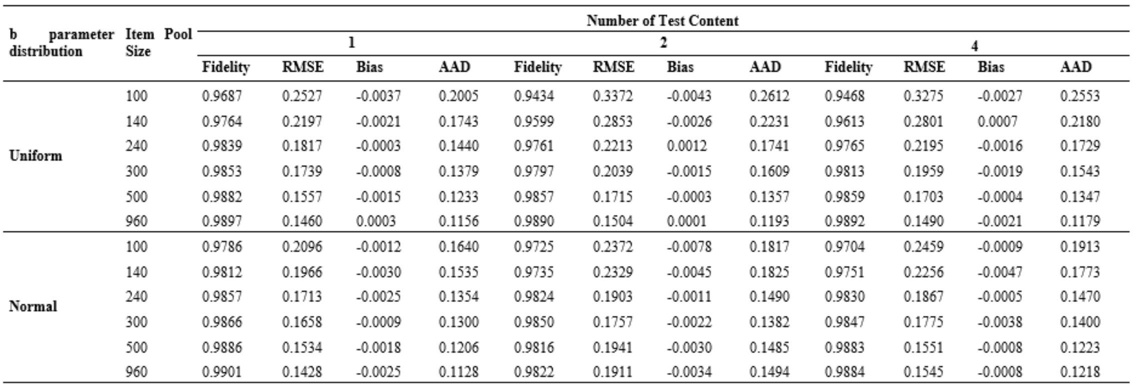
Zhou, X., & Reckase, M. D. (2014). Optimal item pool design for computerized adaptive tests with polytomous items using GPCM. *Psychological Test and Assessment Modeling, 56*(3), 255-274.

Appendix 1: Measurement precision values for each condition in fixed-length termination rule



b parameter distribution	Item Pool Size	Number of Test Content											
		1				2				4			
		Fidelity	RMSE	Bias	AAD	Fidelity	RMSE	Bias	AAD	Fidelity	RMSE	Bias	AAD
Uniform	100	0.9662	0.2625	-0.0022	0.2083	0.9458	0.3302	-0.0023	0.2564	0.9474	0.3255	-0.0029	0.2535
	140	0.9717	0.2405	-0.0031	0.1902	0.9569	0.2954	-0.0032	0.2316	0.9585	0.2899	0.0006	0.2250
	240	0.9800	0.2024	-0.0029	0.1601	0.9697	0.2483	0.0017	0.1938	0.9686	0.2530	0.0015	0.1973
	300	0.9808	0.1986	-0.0012	0.1571	0.9732	0.2339	-0.0002	0.1835	0.9755	0.2238	-0.0013	0.1764
	500	0.9832	0.1857	-0.0026	0.1462	0.9784	0.2103	-0.0009	0.1661	0.9787	0.2087	-0.0010	0.1648
	960	0.9843	0.1796	-0.0010	0.1418	0.9808	0.1983	-0.0008	0.1562	0.9815	0.1949	-0.0016	0.1531
Normal	100	0.9752	0.2250	-0.0025	0.1762	0.9624	0.2764	-0.0043	0.2154	0.9637	0.2717	-0.0008	0.2140
	140	0.9774	0.2152	-0.0040	0.1685	0.9640	0.2707	-0.0041	0.2121	0.9702	0.2466	-0.0038	0.1930
	240	0.9811	0.1971	-0.0018	0.1548	0.9761	0.2209	-0.0022	0.1736	0.9759	0.2218	-0.0023	0.1747
	300	0.9810	0.1977	-0.0007	0.1550	0.9780	0.2123	-0.0026	0.1664	0.9775	0.2148	-0.0009	0.1685
	500	0.9834	0.1844	-0.0013	0.1457	0.9738	0.2314	-0.0040	0.1779	0.9811	0.1971	-0.0024	0.1550
	960	0.9846	0.1776	-0.0025	0.1396	0.9818	0.1933	-0.0030	0.1510	0.9819	0.1929	-0.0019	0.1517

Appendix 2: Measurement precision values for each condition in variable-length termination rule



b parameter distribution	Item Pool Size	Number of Test Content											
		1				2				4			
		Fidelity	RMSE	Bias	AAD	Fidelity	RMSE	Bias	AAD	Fidelity	RMSE	Bias	AAD
Uniform	100	0.9687	0.2527	-0.0037	0.2005	0.9434	0.3372	-0.0043	0.2612	0.9468	0.3275	-0.0027	0.2553
	140	0.9764	0.2197	-0.0021	0.1743	0.9599	0.2853	-0.0026	0.2231	0.9613	0.2801	0.0007	0.2180
	240	0.9839	0.1817	-0.0003	0.1440	0.9761	0.2213	0.0012	0.1741	0.9765	0.2195	-0.0016	0.1729
	300	0.9853	0.1739	-0.0008	0.1379	0.9797	0.2039	-0.0015	0.1609	0.9813	0.1959	-0.0019	0.1543
	500	0.9882	0.1557	-0.0015	0.1233	0.9857	0.1715	-0.0003	0.1357	0.9859	0.1703	-0.0004	0.1347
	960	0.9897	0.1460	0.0003	0.1156	0.9890	0.1504	0.0001	0.1193	0.9892	0.1490	-0.0021	0.1179
Normal	100	0.9786	0.2096	-0.0012	0.1640	0.9725	0.2372	-0.0078	0.1817	0.9704	0.2459	-0.0009	0.1913
	140	0.9812	0.1966	-0.0030	0.1535	0.9735	0.2329	-0.0045	0.1825	0.9751	0.2256	-0.0047	0.1773
	240	0.9857	0.1713	-0.0025	0.1354	0.9824	0.1903	-0.0011	0.1490	0.9830	0.1867	-0.0005	0.1470
	300	0.9866	0.1658	-0.0009	0.1300	0.9850	0.1757	-0.0022	0.1382	0.9847	0.1775	-0.0038	0.1400
	500	0.9886	0.1534	-0.0018	0.1206	0.9816	0.1941	-0.0030	0.1485	0.9883	0.1551	-0.0008	0.1223
	960	0.9901	0.1428	-0.0025	0.1128	0.9822	0.1911	-0.0034	0.1494	0.9884	0.1545	-0.0008	0.1218

Geniş Özet

Bu çalışmada farklı psikometrik özelliklere ve büyüklüklere sahip madde havuzlarının 3 PLM kullanılan bir CAT uygulamasında yetenek kestirimine nasıl etki ettiğinin ve hangi özelliklere sahip madde havuzlarının daha verimli kullanıldığının araştırılması amaçlanmaktadır. Bu amaç doğrultusunda belirlenen alt problemlerin cevaplanması için çalışma kapsamında belirlenen durumlar simüle edilmiştir. Bu nedenle çalışma bir simülasyon çalışmasıdır. Çalışmada veri üretimi ve Bireyselleştirilmiş Bilgisayarlı Test (BBT) simülasyonu için SimulCAT simülasyon yazılımı (Han, 2011) kullanılmıştır. Altı farklı havuz büyüklüğü, iki farklı *b* parametresi dağılımı ve üç farklı kapsam dengelenmesiyle birlikte çalışmaya toplamda 36 (6x2x3) farklı madde havuzu dâhil edilmiştir. Madde havuzu büyüklükleri ilgili alanyazın dikkate alınarak 100, 140, 240, 300, 500 ve 960 olarak belirlenmiştir. Madde havuzlarında yer alan maddelerin *a* parametresi [0.50, 2.00] aralığında ve c parametresi [0.05, 0.20] aralığında tekdüze bir dağılıma sahip olacak şekilde kestirilmiştir. Maddelerin *b* parametresi dağılımı ise normal dağılım olarak adlandırılan madde havuzlarında (*N*(1,1)) olacak şekilde normal dağılıma sahip olurken, tek biçimli dağılım olarak adlandırılan madde havuzlarında [-3, 3] aralığında tek biçimli bir dağılıma sahip olmuştur. Madde havuzlarındaki kapsam sayısı bir, iki ve dört olarak değişmektedir. İki kapsama sahip durumlarda kapsamların dağılımı %50, dört kapsama sahip uygulamalarda kapsamların dağılımı %25 olarak belirlenmiştir. Madde havuzları üretilirken, her bir kapsama ait madde özellikleri bölümünde açıklanan madde dağılımlarına uygun olarak tasarlanmıştır. Simülasyonda ortalaması 1, standart sapması 1 olan (*N*(0,1)) normal dağılımdan 1000 birey oluşturularak katılımcı grubu elde edilmiştir.

Çalışmada ikili puanlanan maddelere dayalı bir BBT uygulaması simüle edilmiştir. Bu nedenle, madde parametrelerinin üretilmesinde 3PLM kullanılmıştır. İlk maddenin seçimi için bireylerin başlangıç yetenek değerleri sıfıra ayarlanmıştır. Yetenek parametresinin kestirilmesinde Beklenen Sonsal Dağılım yöntemi kullanılmıştır. Sonlandırma kuralı olarak sabit test uzunluğu ve değişken uzunluk, seçilmiştir. Sabit test uzunluğu 20 soru olarak belirlenmiş ve değişken uzunluk için theta kestirimindeki standart hata 0.02'den küçük olacak şekilde ayarlanmıştır. Simülasyon, tek bir test zaman aralığı boyunca tek bir BBT oturumuna katılan 1000 katılımcı simüle etmek üzere tasarlanmıştır. Madde seçim yöntemi olarak Maksimum Fisher Bilgisi (MFI) yöntemi kullanılmıştır. Madde kullanım sıklığının kontrol edilebilmesi için Azalarak Kaybolma yöntemi (Han, 2012) (hedef maruz kalma oranı 0.20) seçilmiştir. Kapsam dengelenmesi yöntemi olarak ise Kingsbury ve Zara (1989)'nın Ağırlıklandırma Yöntemi kullanılmıştır. BBT simülasyonu sonucunda uyum katsayısı, RMSE, yanlılık, ortalama mutlak fark (OMF), kestirimin standart hatası (SH), kullanılmayan madde sayısı ve test uzunlukları hesaplanmıştır.

Araştırma sonucunda madde havuzu büyüklüğünün ölçme kesinliğine olan etkisine bakıldığında; her iki sonlandırma kuralında da madde havuzu büyüklüğü arttıkça ölçme kesinliğinin arttığı görülmektedir. Genel olarak, en düşük RMSE, yanlılık ve OMF değerleri en büyük madde havuzlarından elde edilirken, bu göstergelerin en yüksek değerleri ise en küçük madde havuzlarından elde edilmiştir.

Farklı *b* parametresi dağılımına sahip madde havuzları benzer koşullar altında incelendiğinde, az madde sayısına sahip madde havuzlarına ait uyum, RMSE, yanlılık ve OMF değerleri arasındaki farkın, çok maddeye sahip madde havuzlarına ait uyum, RMSE, yanlılık ve OMF değerleri arasındaki farktan daha yüksek olduğu görülmektedir. Yani madde havuzu büyüdükçe *b* parametresinin dağılımının değerler üstündeki etkisinin azaldığı söylenebilir.

Ölçme kesinliği göstergeleri arasında en değişken ve doğrusal olarak değişim göstermeyen yanlılık, madde sayısı arttıkça artıp azalan bir eğilime sahip olmuştur. Sonlandırma kuralı değişen olduğunda ise ölçme kesinliğinin yükseldiği görülmektedir.

Kestirimin standart hatası değerlerinin yetenek düzeyi dağılımlarına bakıldığında; *b* parametresinin normal dağıldığı madde havuzlarında uç değerlerde yüksek, $\theta = 1$ civarında ise en düşük SH değerleri elde edilmiştir. *b* parametreleri $N(1,1)$ şeklinde bir dağılıma sahip olduğu için SH değerlerinin dağılımı bu şekilde olmuştur. Madde havuzunun büyüklüğü arttıkça SH değerleri arasındaki fark azalmıştır. *b* parametresinin tek biçimli dağıldığı madde havuzlarında SH değerlerinin dağılımı küçük madde havuzlarında oldukça farklılık sergilemekte iken büyük madde havuzlarında benzer bir dağılıma sahip olmuştur.

Kapsam sayısı arttıkça SH değerlerinin, her iki *b* dağılımına sahip madde havuzlarında da, arttığı görülmektedir. SH değişiminden en çok 100 ve 140 maddelik madde havuzları etkilenmiştir. *b* parametresinin normal dağılım sergilediği madde havuzlarındaki SH dağılımı kapsam sayısı artsa da benzer dağılıma sahip olmuştur. Ancak *b* parametresinin tek biçimli olarak dağıldığı 100, 140, 240 ve 300 maddelik madde havuzlarında orta yetenek düzeylerinde SH değerlerinin yüksek olduğu görülmüştür.

Farklı sonlandırma kurallarında SH değerlerinin dağılımına bakıldığında; *b* parametresinin normal dağılım sergilediği madde havuzlarında kapsam sayısı artsa da SH değerlerinin çok fazla değişmediği, ancak *b* parametresinin tek biçimli dağıldığı madde havuzlarında kapsam sayısı arttıkça SH değerlerinin küçük madde havuzlarında farklılaştığı görülmektedir. *b* parametresinin tek biçimli dağılım sergilediği bu durum madde havuzu 500'den büyük olduğu durumlarda meydana gelmemiştir. Bu durumda kapsam sayısı 1'den fazla olduğu durumlarda madde havuzu büyüklüğü 500'den fazla olursa SH değerlerinin dağılımının çok değişkenlik göstermediği söylenebilir. Kapsam sayısı 1 olduğunda ise bu durum madde havuzu en az 300 olduğu durumda elde edilmiştir.

Değişen uzunluk sonlandırma kuralına göre; *b* parametresinin normal dağıldığı madde havuzlarında kapsamın 1 olduğu durumlarda test uzunluklarının dağılımı farklı büyüklüklerdeki madde havuzlarında benzer olmuştur. Bu durum *b* parametresinin tek biçimli dağılıma sahip olduğu madde havuzlarında da geçerlidir. Ancak kapsam sayısı arttıkça özellikle küçük madde havuzlarında test uzunluklarının oldukça değişken olduğu görülmüştür. *b* parametresinin normal dağılıma sahip olduğu madde havuzlarında en küçük madde havuzunda yetenek düzeyinin orta düzeyde olduğu kişilerde daha uzun testler uygulanmıştır. Ayrıca tüm koşullarda testlerdeki madde sayısı madde havuzu büyüdükçe artmıştır.

BBT uygulamasında kullanılmayan madde sayılarına bakıldığında; sabit uzunluk sonlandırma kuralının uygulandığı *b* parametresinin dağılımına bakılmaksızın her iki dağılımda da kapsam sayısının 1 olduğu durumda kullanılmayan maddeler yer almıştır. Ancak kapsam sayısı artıkça kullanılmayan madde sayısında azalmalar meydana gelmiştir. Çünkü kapsam dengelemesi yapılırken, madde kullanım sıklığı kontrol yöntemleri de kullanılmıştır. Bu sebeple kullanılmayan madde sayısında azalmalar meydana gelmiştir. Değişen uzunluk sonlandırma kuralının kullanıldığı durumlarda da kapsam sayısı arttıkça kullanılmayan madde sayısında azalmalar meydana gelmiştir. Madde havuzu büyüklüğü arttıkça kullanılmayan maddelerin sayısında artmalar meydana gelmiştir. İki farklı sonlandırma kuralı ile elde edilen sonuçlar karşılaştırıldığında sabit uzunluklu testlerde kullanılmayan madde sayısının daha fazla olduğu görülmektedir.