ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2018; 9(3); 277-294

The Examination of Item Difficulty Distribution, Test Length and Sample Size in Different Ability Distribution

Melek Gülşah ŞAHİN* Yıldız YILDIRIM **

Abstract
This is a post-hoc simulation study which investigates the effect of different item difficulty distributions, sample sizes, and test lengths on measurement precision while estimating the examinee parameters in right and left-skewed distributions. First of all, the examinee parameters were obtained from 20-item real test results for the right-skewed and left-skewed sample groups of 500, 1000, 2500, 5000, and 10000. In the second phase of the study, four different tests were formed according to the *b* parameter values: normal, uniform, left skewed and right skewed distributions. A total of 80 conditions were formed within the scope of this research by selecting 20-item and 30-item condition as the test length variable. In determining the measurement precision, the RMSE and AAD values were calculated. The results were evaluated in terms of the item difficulty distributions, sample sizes, and test lengths. As a result, in right-skewed examinee distribution, the highest measurement precision was obtained at the normal *b* distribution and the lowest measurement precision was obtained at the right skewed *b* distribution. A higher measurement precision was obtained in the 30-item test, however, it was observed that the change in the sample size didn't affect the measurement precision significantly in right-skewed examinee distribution. In the left skewed distribution, the highest measurement precision was obtained at the normal *b* distribution and the lowest measurement precision was obtained at the left-skewed *b* distribution. Also it was observed that the change in the sample size and test length didn't affect the measurement precision significantly in the left-skewed distribution.

Key Words: Item response theory, examinee distribution, item difficulty distribution, sample size, test length.

INTRODUCTION

During the phases of development and scoring process of the tests used to recognize individuals in the fields of Education and Psychology, Classical Test Theory (CTT) and Item Response Theory (IRT) are utilized. These two theories are considered fundamentals in the field of measurement and evaluation. While IRT emerged through the midst of 20th century, the history of CTT dates back to the earlier ages (Crocker & Algina, 1986). IRT is an advantageous and powerful approach in test development, item analysis, and scoring processes (Thompson & Weiss, 2011). Unlike CTT, it is considered that there is a relation between the responses given and the characteristics that the test measures in IRT, and this relation is shown with an increasing function that is named as Item Characteristic Curve (ICC). As IRT does not vary from one group to another, the parameters that determine this curve will remain the same (Lord & Novick, 1968). There are four parameters in the definition of IRT. These are item discrimination parameter (a), item difficulty parameter (b), pseudo guessing parameter (c), and upper asymptote (d). Also, the mathematical equations that describe ICC form IRT models. In addition, the performance of each person who responses the items in the test can be estimated through the instrumentality of the factors named such as characteristics, latent trait or ability (Hambleton, Swaminathan & Rogers, 1991). Another term in the theory is item information function and test information function. The contribution of any item in the scale to the accuracy of measurement done with the whole scale is determined through item information function. Moreover, the test information function is obtained through the total amount of item information function.

*Instructor Dr., Gazi University, Gazi Education Faculty, Ankara-Turkey, e-mail: melekgulsah@gmail.com, ORCID ID: https://orcid.org/0000-0001-5139-9777
** Research Assistant, Adnan Menderes University, Education Faculty, Aydın-Turkey, e-mail: yildiz.yildirim@adu.edu.tr, ORCID ID: https://orcid.org/0000-0001-8434-5062

Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, Cilt 9, Sayı 3, Sonbahar 2018, 277-294.
Journal of Measurement and Evaluation in Education and Psychology, Vol. 9, Issue 3, Autumn 2018, 277-294.
DOI: 10.21031/epod.385000
Received: 28.01.2018
Accepted: 06.08.2018

Item information function and test information function can be obtained independently of sample of individuals. Moreover, these functions are related to standard error of measurement at any ability levels. Due to this features of item information function and test information function is considered as an alternative to reliability and standard error in CTT. The average of test information function at all ability levels means the "reliability" coefficient (marginal reliability) (Hambleton & Swainathan, 1985).

Unidimensionality, local independence and normality assumptions are found in the unidimension and parametric models of IRT. Unidimensionality assumption is based on the statistical independence among items (Crocker & Algina, 1986) and test items measure only one ability (Hambleton et al., 1991). Local independence assumption is related to unidimensionality and it means that, when the abilities influencing the test performance of the individuals are at the same level, individuals' responses to any pair of items are statistically independent from the responses to any other test items. Although unidimensionality and local independence are different terms, when the test ensures its unidimensionality, it means that the local independence assumption is obtained (Hambleton et al., 1991).

The characteristic features of IRT has improved test development, test bias identification, test equating and the limitations have been removed in these conditions (Hambleton & Swaminathan, 1985). Thanks to the advantages of IRT, this theory has been preferred in the examinations especially like PISA (The OECD Programme for International Student Assessment) and TIMSS (The Trends in International Mathematics and Science Study) (Martin, Mulis & Hooper, 2016; OECD, 2017). In addition, it is seen in many national and international research that test results are evaluated within the context of IRT (Ackermann, 1994; Bhakta, Thennant, Horton, Lawton & Andrich, 2005; Çelen & Aybek, 2013; İlhan, 2016). The exams used in education are prepared for many different purposes, and these exams are extremely important for individuals. These purposes can include student selection and placement, proficiency, diagnostic tests etc. These tests will have various psychometric characteristics depending on the purpose of development, the characteristics of individuals or the number of individuals taking the test. For example, if the number of students are more but the number of the students to be selected according to the results is less, the test can be expected to be difficult. However, if the test is to be developed to diagnose the existing knowledge (not to select and place), the test is expected to be easier than selection and placement tests and to consist of items with moderate difficulty, if possible. It is more important here to identify how the validity and reliability will be affected in the tests that have different item difficulty index. In addition, how the ability distribution of the individuals that take the test affect the validity and reliability should also be identified. In this study, based on the results of a national exam, the effect of test length and sample size for different ability distributions in the tests that have different *b* parameters within ability parameter estimation on measurement precision was analyzed.

In the literature, there are studies that analyze the effect of sample size on measurement precision in various models and items with different scores in the item response theory (Boughton, Klinger & Gierl, 2001; Cheng & Yuan, 2010; De Ayala & Bolesta, 1999; DeMars, 2002; DeMars, 2003; Montgomery & Skorupski, 2012; Preston & Reise, 2014). In addition to these, there are studies which consist at least two of sample size, test length and ability distribution type conditions. (Ankenmann ve Stone, 1992; Baker, 1998; Guyer ve Thompson, 2011; Hulin, Lissak ve Drasgow, 1982; Kieftenbeld ve Natesan, 2012; Lautenschlager, Meade ve Kim, 2006; Preinerstorfer ve Formann, 2012; Roberts ve Laughlin, 1996; Seong, Kim ve Cohen, 1997; Stone, 1992; Swaminathan ve Gifford; 1979; Wang ve Cheng, 2005; Wollack, Bolt, Cohen ve Lee, 2002). Furthermore, while there are studies that *a* parameter is obtained within different ranges and that analyze its impact on measurement precision (DeMars, 2003; Preston & Reise, 2014; Reise & Yu, 1990), fewer studies examine *b* parameters' impact on measurement precision. Some studies related to this study are summarized as follows.

Lautenschlager et al. (2006), in a post-hoc simulation study within graded response model (GRM), examined the effect of 7 different sample sizes (75, 150, 200, 300, 500, 1000 and 2000 individual), four different test lengths (5, 10, 15 and 20 items), and three different sample distributions (normal,

skewed and uniform) on ability and item parameter estimation. The researchers used maximum posteriori (MAP) estimation method in the ability parameter estimation. In the study, the results showed that sample size does not change the root mean squared error (RMSE) values but RMSE values decreased when the test length increases. Ankenmann and Stone (1992) carried out a post-hoc simulation study using three different test lengths (5, 10, and 20 items), with a sample size of 125, 150, 500 for one-parameter GRM and with a sample size of 250, 500, and 1000 for 2-parameter GRM, they analyzed how ability estimation was affected. The researchers that used marginal maximum likelihood (MML) in parameter estimation used MULTILOG Program. As a result, it was concluded that sample size did not have an important effect on ability parameter estimation. In addition, it was found that the longer the test length is, the more precise the measurement in ability estimation. Kieftenbeld and Natesan (2012) conducted another post-hoc simulation in their study using a four different test lengths (5, 10, 15, and 20 items), five different sample sizes (75, 150, 300, 500, and 1000 individuals) and three different ability distribution types (normal, uniform, and skewed), and they analyzed the effect of these conditions on ability and item parameter. In the study, MML and Markov Chain Monte Carlo (MCMC) methods were used for estimation. They conducted the study within the context of GRM and estimated the parameters using MULTILOG program. The results of the study revealed that test length described the highest variance in RMSE whereas sample size described a less amount of the variance. Preinerstorfer and Formann (2012) analyzed the effect of two different sub-groups (1 and 2 sub-group), homogeneity and heterogeneity of the groups, four different test lengths (10, 15, 25 and 40 items) and three sample sizes (500, 1000, and 2500) on measurement precision in parameter estimation using mixed Rasch model. As a result, it was found that as sample size and test length increased, so did the measurement precision.

In the literature, for the models related to polytomous items and Rasch model, there are some studies that analyze the effect of sample size and/or test length on measurement precision, and some other similar studies with logistic models related to dichotomous items. For example, Swaminathan and Gifford (1979) analyzed the effect of ability and item parameter estimation on measurement precision using Urry and MLE methods. They used different test lengths (10, 15, 20, and 80), different sample sizes (50, 200, and 1000), and different ability distribution types (normal, uniform, and skewed) within 3PL model. As a result, they stated that when the sample size and test length increased, so did the measurement precision within ability parameter, and there was a little effect of sample size on measurement precision. Hulin et al. (1982) carried out a Monte-Carlo study using 2PL and 3PL models and analyzed the effect of different sample sizes (200, 500, and 1000), different test lengths (15, 30, and 60) on measurement precision within item and ability parameter estimation. The result of the study revealed that the accuracy of ability estimation in 3PL is less in small samples and small lengths. In addition, it was found that the sample size in 30 and 60 item tests in 3PL model did not affect RMSE and correlation values much. Stone (1992) analyzed the effect of different sample sizes (250, 500, and 1000), different test lengths (10, 20, and 40) and different distribution types (normal, skewed, and platykurtic) in 2PL model on measurement precision within parameter estimation. The result of the study revealed that the most significant condition that affected measurement precision was test length within ability parameter estimation (especially among extreme ability parameters). In addition, it was found that when the test length gets longer, error of estimation decreased significantly. Furthermore, they also found that the increase in the sample size did not reduce the deviation. Stone also analyzed the measurement precision within item level and the effect of research conditions when *b* parameter was in different levels (average (0, 02), easy (-2, 18), difficult (1, 82)) on measurement precision. In this context, it was found that when the item difficulty was average, lower RMSE values were achieved within item parameter estimation, and the highest RMSE values were seen in easy items. Cheng and Yuan (2010) aimed to correct the standard error of ability estimation using MLE method within 2PL model. These researchers, who analyzed the effect of sample size on standard error, determined the sample size as 200 and 2000. It was found that the increase in the sample size did not affect the standard error significantly.

Finally, some studies that analyze the effect of sample size and test length on measurement precision are summarized below. Köse (2010) aimed to analyze the effect of different sample sizes (500, 1000,

and 1500) and different test lengths (12 and 24) on item and ability parameter estimation and model data fit in unidimensional (2PL) and multidimensional models. The results of the study reveal that sample size in ability parameter estimation did not have a significant effect on both unidimensional and multidimensional models. In addition, Köse stated that, based on RMSD values, the increase in the number of items in ability parameter estimation caused less defective results. Koğar (2015) carried out a Monte Carlo study using unidimensional, unidimensional non-parametric and multi-dimensional IRT models and analyzed the effect of different sample sizes (100, 500, 1000, and 5000), different test lengths (5, 15, and 25) and different inter-dimensional correlation values (0,00, 0,25, and 0,50) on item parameter estimation and model fit. The results suggested that, in unidimensional and multidimensional models, in order for the item parameter estimation to be more accurate, the sample size and test length should be greater.

In the literature, the studies usually focus on analyzing the effect of some variables such as sample size, test length, and item discrimination index on measurement precision within ability parameter estimation. Different from many studies, this study investigated how the measurement precision of the ability parameter estimation is affected by different *b* parameter distributions (normal, uniform, right-skewed, and left-skewed), in addition to analyzing the effect of sample size and test length in left and right skewed ability distributions.

Purpose of the Study

This study aims to analyze the effect of different *b* parameter distributions, test lengths, sample sizes on measurement precision of ability parameter estimation in right skewed and left-skewed ability distributions. It was found that literature generally focuses on different conditions that affect measurement precision within ability parameter estimation. As stated in the introduction part of this study, the studies usually analyze the effect of sample size and test length on measurement precision. However, no studies were found in literature that analyze the effect of different *b* parameter distributions on measurement precision in the groups that have different ability distributions, different test lengths and sample sizes. Production of four different tests based on different item difficulty distributions is considered important. The problem of the study is "what is the effect of different item difficulty distributions, sample sizes, and test lengths in right-skewed and left-skewed ability distributions on measurement precision of ability parameter estimation?"

Sub-problems of the study are as follows:

1. What is the effect of different test lengths, sample sizes, item difficulty distributions within right-skewed ability distribution on measurement precision of ability parameter estimation?

2. What is the effect of different test lengths, sample sizes, item difficulty distributions within left-skewed ability distribution on measurement precision of ability parameter estimation?

METHOD

Data Production

Obtaining Ability Parameter Values

In this post-hoc simulation study, real data were used to collect ability parameters. The real data were obtained from the 20-items mathematics subtest of Placement Test (Seviye Belirleme Sınavı-SBS) applied in 2012. This placement test was used to select students who will continue high school education. In the study, totally five sample sizes (500, 1000, 2500, 5000, and 10000) were chosen from the data set. Previous studies in the literature (Ankenmann & Stone, 1992; Baker, 1998; DeMars, 2002; Guyer & Thompson, 2011; Hulin et al., 1982; Kieftenbeld & Natesan, 2012; Lautenschlager et al., 2006; Montgomery & Skourpski, 2012; Preinerstorfer & Formann, 2012; Preston & Reise; 2014; Reise & Yu, 1990; Roberts & Laughlin, 1996; Seong et al., 1997; Stone, 1992; Swaminathan & Gifford, 1979; Thissen & Wainer, 1982; Wang & Cheng, 2005; Wollack et

al., 2002, Yavuz & Hambleton, 2016) were utilized while choosing the sample size. For each sample size chosen for obtaining the ability parameters, both right-skewed and left-skewed ability, distributions were chosen from the real data. During the selection of right and left-skewed distributions for each sample size for the right-skewed distribution, SBS data, which is originally a right-skewed data set (coefficient of skewness=1,05), was done randomly. For the left-skewed data sets, similar to the study of Doğan and Tezbaşaran (2003), intended sample distribution was achieved through purposive sampling, and the groups whose coefficient of skewness is ≈-1,00 were chosen for all sample sizes.

Similar to the coefficient of skewness values used in Doğan and Tezbaşaran (2003), Bahry (2012) and Sen (2014), it was determined the coefficient of skewness as +1,00 in this study. For the left-skewed distribution, Doğan & Tezbaşaran (2003) and Bıkmaz Bilgen & Doğan (2017) used a -1,00 coefficient of skewness in their studies. After these groups were chosen from the areal data, maximum likelihood estimation method was used in MULTILOG 7.03 program (Thissen, Chen & Bock, 2003) and the groups' ability parameters were estimated with 25 replications, and this post-hoc simulation study was completed.

Simulation of Item Parameters

In the second step of the study, different four tests were created which have different *b* parameters: tests with normal distribution, uniform distribution, right-skewed and left-skewed distribution. The statistics used in test development were determined according to the values and suggestions within the studies in the literature (Ankenmann & Stone, 1992; Baker, 1998; Bahry, 2012; De Ayala & Sava-Bolesta,1999; DeMars, 2002; DeMars, 2003; Dolma, 2009; Fotaris, Mastoras, Mavridis & Manitsaris, 2010; Han, 2012; Hulin et al., 1982; Kieftenbeld & Natesan, 2012; Montgomery & Skourpski, 2012; Preston & Reise; 2014; Reise & Yu, 1990; Seong et al., 1997; Stone, 1992; Swaminathan & Gifford, 1979). In accordance with these studies, *a* parameter value was determined as min=0,5 and max=2 in the simulation of item parameters, and *c* parameter value was determined as min=0 and max=0,05. Four different item difficulty distribution were created for left-skewed *b* parameter α=8; β=2; for right-skewed *b* parameter distribution α=2; β=8; for uniform *b* parameter distribution min=-3; max=+3 and for normal *b* parameter distribution average=0 and sd=1 values were used. For the test length variable of the study, two different conditions with 20 and 30 items were determined. The reason why the test length was determined as 20 and 30 items is that these test lengths are mainly used in national exams and the studies in the literature use similar test lengths (Ankenmann & Stone, 1992; Baker, 1998; Boughton et al., 2001; Craig & Kaiser, 2003; DeMars, 2003; Fotaris et al., 2010; Guyer & Thompson, 2011; Hulin et al., 1982; Kieftenbeld & Natesan, 2012; Lautenschlager et al., 2006; Roberts & Laughlin, 1996; Seong et al., 1997; Stone, 1992; Swaminathan & Gifford, 1979; Wang & Cheng, 2005; Wollack et al., 2002, Yavuz & Hambleton, 2016). 80 conditions (2 ability distribution, x5 sample size, x4 *b* parameter distribution, x2 test length) dealed within the scope of the study were created via WinGen 3 program (Han, 2007; Han & Hambleton, 2007) after 25 replications. Within the scope of the study, the reason why 25 replications were made is that it is a sufficient number in the elimination of sample bias (Harwell, Stone, Hsu & Kirisci, 1996).

Data Analysis

During data analysis process, firstly ability parameter estimation produced data were done through MULTILOG 7.03 and 2000 times (80 conditions x 25 replication) based on MLE method. Then the estimated measurement precision of ability parameter was analyzed as parameter recovery studies in IRT generally use measurement precision calculation. To analyze measurement precision, RMSE and "average absolute deviation (AAD)" values were calculated. RMSE and AAD values were calculated after each replication and compared to the number of replications, then the average score was reported and discussed. To calculate these values, the following formulas were used:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(\hat{\theta_j} - \theta_{Tj})^2}{N}}$$

$$AAD = \frac{\sum_{i=1}^{N}\left|\widehat{\theta_J} - \theta_{T_j}\right|}{N}$$

In these formulas, θ_{T_j} j. means actual ability parameter for the individual; $\widehat{\theta_J}$ j refers to ability parameter estimated for the individual and N describes the sample size. When RMSE and AAD values get closer to 0, the measurement precision increases. Thus, the accuracy of parameter estimation also increases. In addition, some interpretations were made according to the criterion that RMSE value is less than 0,10 (DeMars, 2003; Sen, Cohen & Kim, 2015; Tate, 2000).

RESULTS

This part represents the findings within the context of sub-problems of the study.

1. Sub-problem: What is the effect of different test lengths, sample sizes, item difficulty distributions within right-skewed ability distribution on measurement precision of ability parameter estimation?

All the RMSE and AAD values from analysis done for right-skewed ability distribution are shown in Table 1.

Table 1. RMSE and AAD Values in Right-Skewed Ability Distribution in Relation to Test Conditions

Right-Skewed Ability Distribution		Item Difficulty Parameter Distribution							
		Normal		Uniform		Left-Skewed		Right-Skewed	
Test Lengths	Sample Sizes	RMSE	AAD	RMSE	AAD	RMSE	AAD	RMSE	AAD
20	500	0,080	0,317	0,112	0,460	0,144	0,562	0,235	1,108
	1000	0,080	0,320	0,115	0,469	0,150	0,587	0,232	1,087
	2500	0,079	0,315	0,112	0,459	0,149	0,583	0,231	1,089
	5000	0,079	0,314	0,112	0,458	0,148	0,581	0,232	1,091
	10000	0,079	0,315	0,112	0,460	0,148	0,580	0,232	1,090
30	500	0,070	0,275	0,101	0,411	0,156	0,637	0,231	1,101
	1000	0,071	0,282	0,102	0,419	0,163	0,665	0,228	1,078
	2500	0,070	0,279	0,100	0,408	0,161	0,663	0,228	1,081
	5000	0,070	0,278	0,100	0,408	0,161	0,663	0,228	1,082
	10000	0,070	0,280	0,100	0,411	0,161	0,661	0,228	1.082

In Table 1, RMSE and AAD values, which were used to determine the measurement precision for 40 conditions within right-skewed distribution, are represented. In this sub-problem, the variation of RMSE and AAD values (in different *b* parameter distributions and sample size for 20 and 30 test items within the context of right-skewed ability distribution) is shown in Figure 1 and the figures are discussed with Table 1.

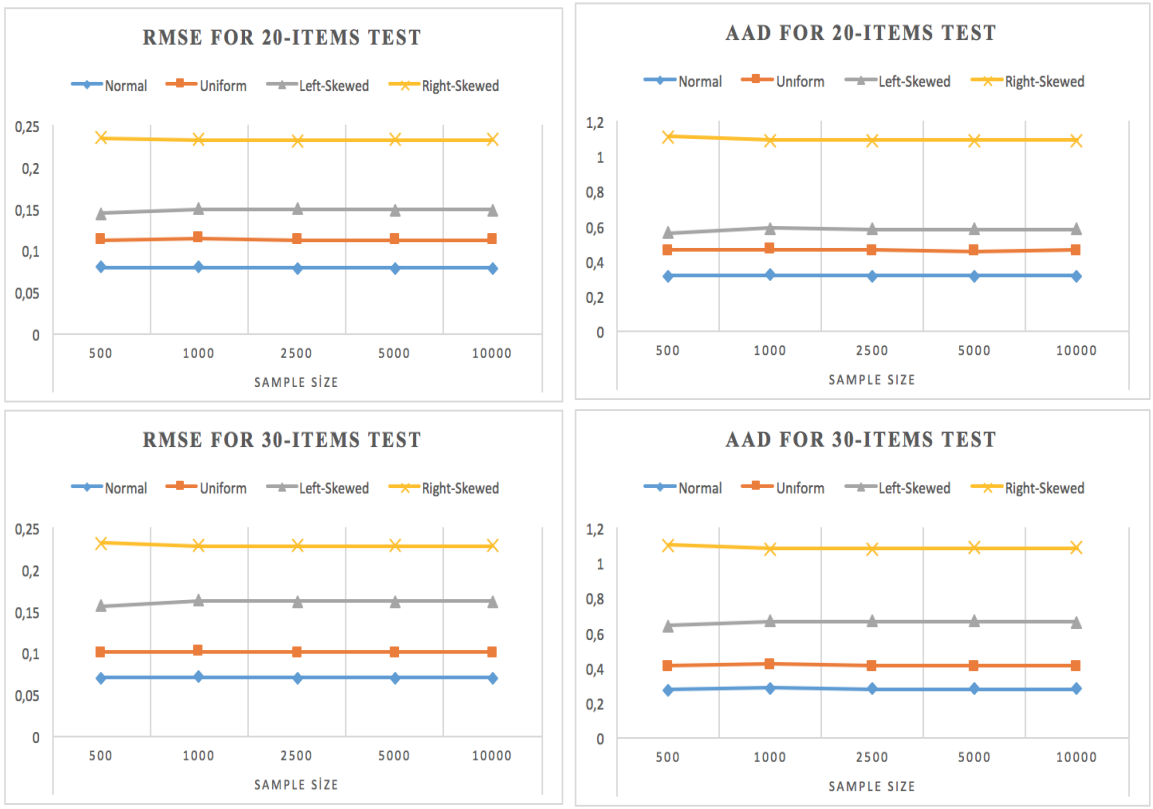

Figure 1. Graphics in Relation to RMSE and AAD within the Context of Test Length for Right-Skewed Ability Distribution.

When Figure 1 and Table 1 are analyzed, within all sample sizes (500, 1000, 2500, 5000, and 10000) that has right-skewed ability distribution, when *b* parameter distribution is normal, it can be seen that the lowest RMSE and AAD values were obtained for both 20-item test and 30-item test. These RMSE and AAD values are followed by uniform and left-skewed distribution for all sample sizes respectively. However, the highest RMSE and AAD values were obtained from the distribution in which *b* parameter has right-skewed distribution. Based on these values of RMSE and AAD statistics, it can be stated that, within all sample sizes, the measurement precision is the highest when *b* parameter has a normal distribution and the lowest when it has right-skewed distribution, and the second highest measurement precision distribution type is the uniform distribution. In addition, sample size did not have much effect on RMSE and AAD values within ability parameter estimation within different *b* parameter distribution and test lengths for right-skewed ability parameter. This result can be seen in Figure 1 and Table 1. In other words, sample size did not have a significant effect on measurement precision within ability parameter estimation.

With reference to the values in Table 1, the variation of RMSE and AAD values within different *b* parameter distributions and test lengths (individually for each sample size) is shown in Figure 2 and the figures are discussed with Table 1.

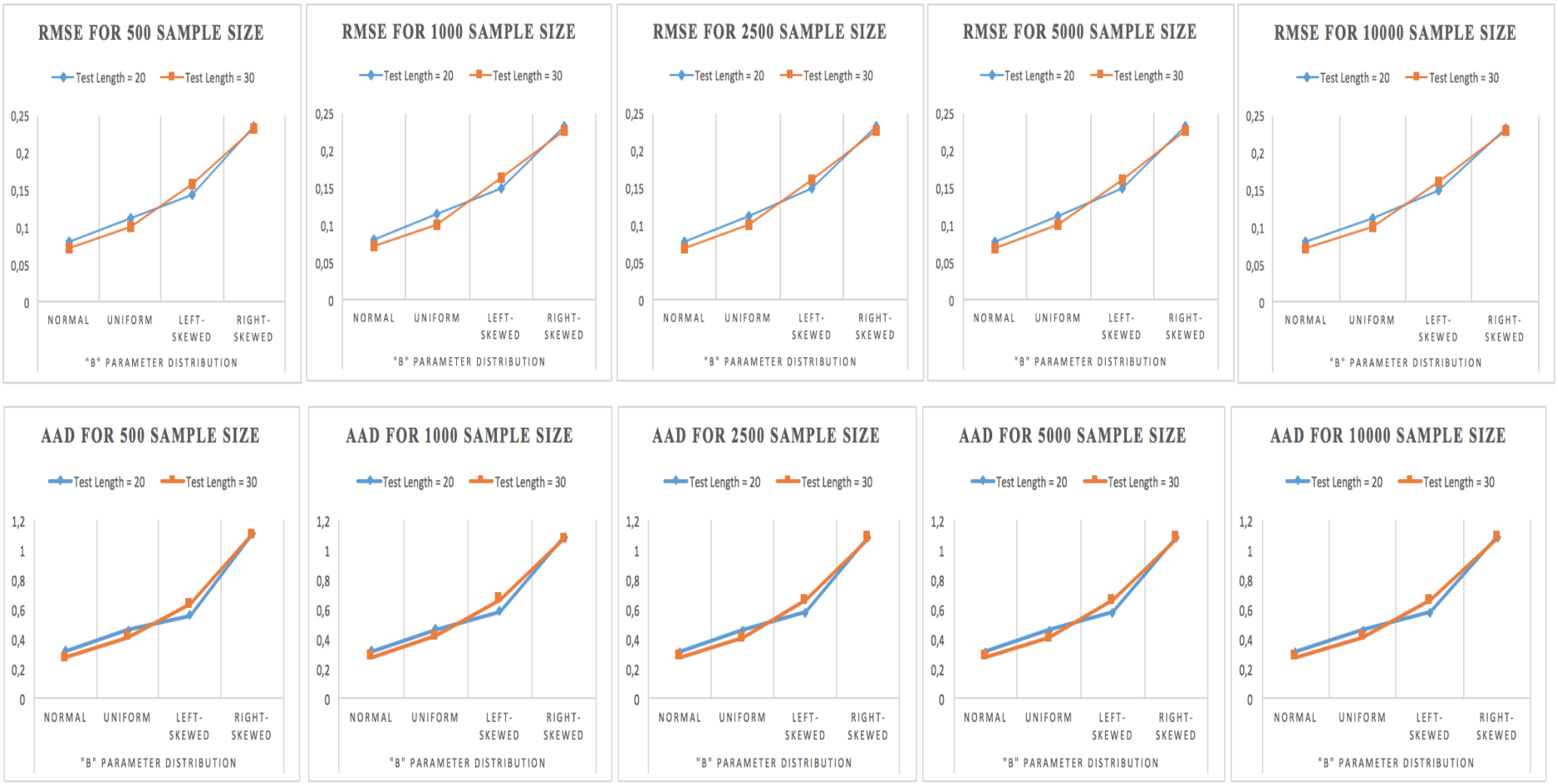

Figure 2. Graphics in Relation to RMSE and AAD Values within the Context of Sample Size for Right-Skewed Ability Distribution.

When Figure 2 and Table 1 is examined, when b distribution is normal, it can be seen that the lowest RMSE and AAD values were obtained in 30-items test. Higher RMSE and AAD values were obtained for 20 items within each sample size than the values within 30-item test. When item difficulty parameter has uniform and right-skewed distribution, for all sample sizes, the lowest RMSE and AAD values, similar to the distribution in normal item difficulty, was seen within 30-item test. Accordingly, it can be said that, in the tests that have normal, uniform, and right-skewed *b* parameter distribution, for all sample sizes, when the test length increases, the measurement precision also increases. However, for the left-skewed *b* parameter distribution, when all sample sizes are considered, the lowest RMSE and AAD values were obtained from 20-item test. It was different from the other item difficulty distributions. This may be because of the increase in the number of items with high item difficulty. Overall, when the test length increases, RMSE and AAD values decrease; and hereby measurement precision increases. When the values for right-skewed ability parameter are analyzed, it was found that, for all *b* parameter distributions, the values obtained from different test lengths were more or less the same. However, it was also seen that, in contrast with sample size, the values varied when test length changes. In conclusion, it can be stated that, based RMSE<0,10 on the criteria that Tate (2000), DeMars (2003) and Sen et al. (2015) used, all test lengths and sample sizes were convenient when the *b* parameter distribution is normal. However, in other *b* parameter distributions, all of test lengths and sample sizes were not found appropriate based on the criterion.

2. Sub-problem: What is the effect of different test lengths, sample sizes, item difficulty distributions within left-skewed ability distribution on measurement precision of ability parameter estimation?

All RMSE and AAD values obtained from the whole analysis for left-skewed ability distribution are shown in Table 2.

Table 2. RMSE and AAD Values in Left-Skewed Ability Distribution in Relation to Test Conditions

Left-Skewed Ability Distribution		Item Difficulty Parameter Distribution							
		Normal		Uniform		Left-Skewed		Right-Skewed	
Test Length	Sample Size	RMSE	AAD	RMSE	AAD	RMSE	AAD	RMSE	AAD
20	500	0,079	0,324	0,137	0,610	0,246	1,166	0,149	0,652
	1000	0,079	0,326	0,136	0,610	0,248	1,183	0,147	0,656
	2500	0,079	0,326	0,138	0,616	0,250	1,191	0,146	0,638
	5000	0,079	0,328	0,137	0,611	0,250	1,192	0,146	0,640
	10000	0,079	0,327	0,138	0,617	0,250	1,191	0,146	0,639
30	500	0,078	0,322	0,137	0,610	0,248	1,176	0,150	0,656
	1000	0,079	0,327	0,137	0,615	0,249	1,184	0,147	0,643
	2500	0,079	0,327	0,135	0,604	0,250	1,191	0,146	0,641
	5000	0,079	0,327	0,138	0,617	0,249	1,189	0,146	0,639
	10000	0,079	0,326	0,138	0,617	0,250	1,190	0,146	0,639

In Table 2, RMSE and AAD values, which were used to determine the measurement precision for 40 conditions within left-skewed distribution, are represented. In the second sub-problem, the variation of RMSE and AAD values (in different b parameter distributions and sample size for 20 and 30 test items within the context of left-skewed ability distribution) is shown in Figure 3 and the figures are discussed with Table 2.

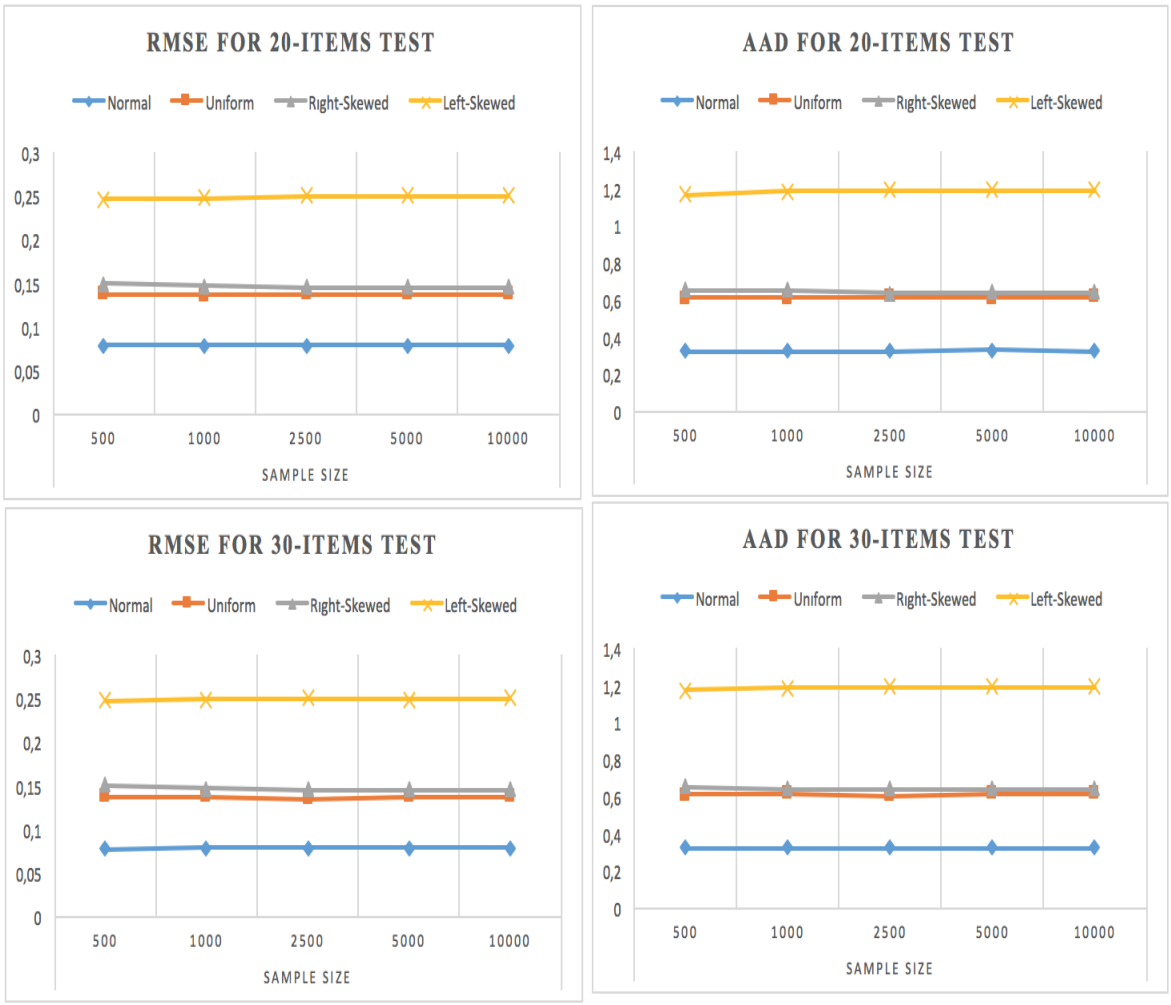

Figure 3. Graphics in Relation to RMSE and AAD Values within the Context of Test Length for Left-Skewed Ability Distribution.

When Figure 3 and Table 2 is examined, when b distribution is normal, within all sample sizes that have left-skewed ability distribution, it can be seen that the lowest RMSE and AAD values were obtained for both 20-items test and 30-item tests. These values are followed by uniform b distribution and right-skewed distribution respectively. The highest RMSE and AAD values were obtained from the distribution in which *b* parameter has left-skewed distribution. Based on these values of RMSE and AAD statistics, it can be stated that, within all sample sizes, the measurement precision is the highest when *b* parameter has a normal distribution and the lowest when it has left-skewed distribution, and the second highest measurement precision distribution type is the uniform distribution. In addition, sample size did not have much effect on RMSE and AAD values within ability parameter estimation within different *b* parameter distribution and test lengths for left-skewed ability parameter distribution. This result can be seen in Figure 3 and Table 2. In other words, sample size did not have a significant effect on measurement precision within ability parameter estimation. The variation of RMSE and AAD values within different *b* parameter distributions and test lengths (individually for each sample size) is shown in Figure 4 and the figures are discussed with Table 2.

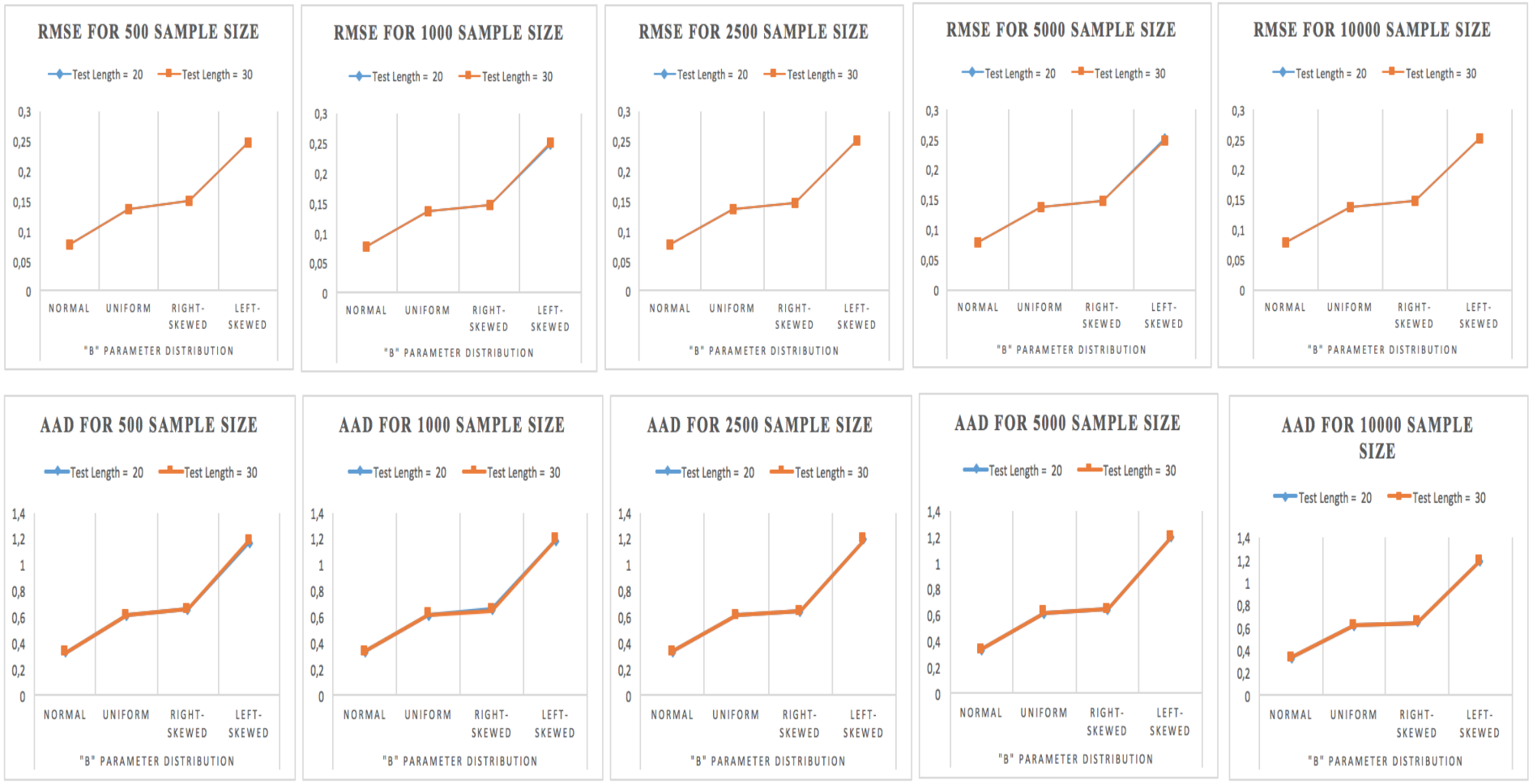

Figure 4. Graphics in Relation to RMSE and AAD Values within the Context of Sample Size for Left-Skewed Ability Distributio

When Figure 4 and Table 2 is analyzed, for left-skewed ability parameter distribution, it was seen that RMSE and AAD values are similar in both 20-item and 30 item within all item difficulty parameter distributions and sample sizes. Accordingly, it can be said that, within all sample sizes and item difficulty parameter distributions, measurement precision does not change significantly although the test length increases. In conclusion, it can be stated that, based RMSE<0,10 on the criteria that Tate (2000), DeMars (2003) and Sen et al. (2015) used, all test lengths and sample sizes were convenient when the *b* parameter distribution is normal. However, in other *b* parameter distributions, all of test lengths and sample sizes were not found appropriate based on the criterion.

DISCUSSION and CONCLUSION

In this study, measurement precision of ability parameter estimation obtained from the conditions that are generated from two different ability distribution, five different sample size, four different *b* parameter distribution, and two different test length is analyzed. The ability parameter values were estimated according to the conditions addressed by the data from a national exam. To determine the test lengths, the average test lengths of national exams were considered. To create the tests, it is considered that the conditions in which *b* parameter comprised of normal, uniform, right-skewed, and left-skewed distributions.

When the results for right-skewed ability distribution are examined, it is seen that, when the sample size of each test that has different *b* parameter distribution increases, RMSE and AAD values that are measured for measurement precision do not change significantly. When the effect of sample size change for 20-items and 30-items tests is examined, it is seen that RMSE and AAD values decrease when sample size increases. However, when the conditions in which sample size and test length has different *b* parameter distributions, the best results were obtained when *b* parameter has normal distributions. This condition is followed by the condition which *b* parameter has uniform distribution. In the conditions that has uniform distribution, similar to other conditions, there is not a significant effect of different sample sizes on measurement precision. When *b* parameter had left-skewed distribution, RMSE and AAD values did not vary much in different sample sizes but they decreased when test length increased. Lower RMSE and AAD values were obtained for 30 items than 20-items test when *b* parameter distribution had right-skewed. In addition, it can be stated that, when sample size increases, RMSE and AAD values do not vary significantly but the difference between 500 and 1000 individuals are higher than other sample sizes. In right-skewed b distribution, RMSE and AAD values were higher than other b distributions. Similarly, Stone (1992) compared normal ability distribution for easy items and right-skewed ability distribution and found that right-skewed ability distribution (such conditions as 20 items and 500-1000 sample size) had lower measurement precision values than normal ability distribution.

When left-skewed ability distribution was examined, it is seen that, when sample size for each test that has different *b* parameter increased, RMSE and AAD values did not have significant change. When the effect of test length was analyzed, it was found that in the group that had left-skewed ability parameter, the increase of the test length did not affect measurement precision in general. When the effect of item difficulty parameter was examined, it was found that the lowest RMSE and AAD values were obtained when *b* parameter had normal distribution. This distribution was followed by uniform *b* parameter distribution (relevant for both test lengths and all sample sizes). It was found that by achieving the highest RMSE and AAD values in left-skewed *b* parameter distribution and measurement precision was the lowest for these values.

The overall results of the study showed that, within both left-skewed and right-skewed ability parameter distribution, when the sample size within each *b* parameter distribution types increases, no significant change was observed in measurement precision. In the literature, some studies show the same results for similar conditions. Hulin et al. (1982) and Swaminathan and Gifford (1979), for example, stated that sample size does not have a significant effect on RMSE and correlation values.

Stone (1992) and Cheng and Yuan (2010), within two-parameter logistic model, found that sample size does not affect error significantly within the estimation of ability parameters.

The result of the study showed that the best estimations for both left-skewed and right-skewed ability parameter distribution was observed in condition which b distribution was normal. Stone (1992) stated that, within right-skewed and normal ability parameter distribution, the best estimations appear in condition that the item difficulty is medium. In addition, he added that the worst estimations appear within easy items. Similarly, in this study, for right-skewed ability parameter distribution, the most defective estimations are made when *b* parameter distribution is right-skewed. Wollack et al. (2002) stated that parameter recovery is best done with the medium-difficulty items and worst done with extreme (easy or difficult) items. Similarly, in this study, Yen (1987) analyzed the conditions in which item difficulty is easy, average and difficult, and worked with 20-items test length, normal ability distribution and with the sample size of 1000. The results of his study revealed that the highest measurement precision was obtained from medium-difficulty items.

Findings about the effect of test length show that, within right-skewed ability distribution and other conditions (normal, uniform, and right-skewed) except for left-skewed item difficulty distribution, measurement precision increases when test length increases. In the literature, there are similar studies in accordance with the relevant results of dichotomous models and polytomous models (Ankenmann & Stone, 1992; Boughton et al., 2001; Hulin et al., 1982; Kieftenbeld & Natesan, 2012; Lautenschlager et al., 2006; Preinerstorfer & Formann, 2012; Roberts & Laughlin, 1996; Seong et al., 1997; Stone, 1992; Swaminathan & Gifford, 1979). For 3PL of dichotomous models Swaminathan and Gifford (1979), Hulin et al. (1982) and for 2PL Stone (1992) identified that measurement precision increase when test length increases. For left-skewed ability distribution, no effect of test length was observed. In the literature, there are studies which the ability estimation of test length do not affect measurement precision (Wollack & Cohen, 1998; Wollack et al., 2002). Wollack et al. (2002) had similar results to this study. They found that the increase of test items from 20 to 30 does not develop $P_{ik}(\theta j)$ estimation.

In this study, in accordance with the results obtained from the individuals who have right-skewed ability parameter, it can be suggested that test developers should ensure that number of items is higher when *b* parameters are distributed normal, uniform or right-skewed, and ensure that number of items is lower when *b* parameters have left-skewed as long as it does not decrease content validity. In addition, as measurement precision will be higher when *b* parameter distribution is normal (independently from ability parameter), it is suggested that *b* parameters in the test should have normal distribution as long as it is relevant with the purpose. In other words, when most of the items have a medium-difficulty level, it would be more appropriate in accordance with the results if difficult and easy items are fewer. Another suggestion for the test developers is that most of the test items should not be very difficult (when *b* parameter distribution is left-skewed) or very easy (when *b* parameter distribution is right-skewed). Because within this kind of *b* parameter distributions, measurement precision may be lower when compared to normal and uniform distribution.

In this study, right-skewed and left-skewed ability parameters were produced from the real data, and conditions were created with reference to different sample size, different *b* parameter distributions and different test lengths. Other researchers can conduct some other studies in other conditions that have estimation method, model, and number of categories for polytomous items, number of replication, estimation program etc. rather than sample size and test length. In addition, they can research the effect of different *b* parameter distributions on measurement precision when ability parameters have normal and uniform distribution. While this study was conducted for dichotomous data, other studies can be conducted for polytomous. Although this study was done using 3-parameter logistic model, other researchers can use other models. In conclusion, while this study analyzed measurement precision within ability parameter estimation, some other studies, within same conditions, can analyze the change of measurement precision within item parameter estimation.

REFERENCES

Ackerman, T. A. (1994). Using multidimensional item response theory to understand what items and tests are measuring. *Applied Measurement in Education*, *7*(4), 255-278. Doi: 10.1207/s15324818ame0704_1

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*, 561-573. doi: 10.1007/BF02293814

Ankenmann, R. D., & Stone, C. A. (1992, April). *A Monte Carlo study of marginal maximum likelihood parameter estimates for the graded model*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Bahry, L. M. (2012). *Polytomous item response theory parameter recovery: An investigation of non-normal distributions and small sample size* (Unpublished Master Thesis, University of Alberta Department of Educational Psychology, Edmonton). Retrieved from https://era.library.ualberta.ca/items/55cebca1-82a2-44b5-ab78-aad933bbf147.

Baker, F. B. (1998). An investigation of the item parameter recovery characteristics of a Gibbs sampling procedure. *Applied Psychological Measurement*, *22*(2), 153-169. doi: 10.1177/01466216980222005

Bhakta, B., Tennant, A., Horton, M., Lawton, G., & Andrich, D. (2005). Using item response theory to explore the psychometric properties of extended matching questions examination in undergraduate medical education. *BMC Medical Education*, *5*(1), 9. doi: 10.1186/1472-6920-5-9

Bıkmaz Bilgen, Ö., & Doğan, N. (2017). Çok kategorili parametrik ve parametrik olmayan madde tepki kuramı modellerinin karşılaştırılması. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *8*(4), 354-372. doi: 10.21031/epod.346650

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*(1), 29-51. doi: 10.1007/BF02291411

Boughton, K. A., Klinger, D. A., & Gierl, M. J. (2001, April). *Effects of random rater error on parameter recovery of the generalized partial credit model and graded response model*. Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle, WA.

Cheng, Y., & Yuan, K. H. (2010). The impact of fallible item parameter estimates on latent trait recovery. *Psychometrika*, *75*(2), 280-291. doi: 10.1007/s11336-009-9144-x

Craig, S. B., & Kaiser, R. B. (2003). Applying item response theory to multisource performance ratings: What are the consequences of violating the independent observations assumption? *Organizational Research Methods*, *6*(1), 44-60. doi: 10.1177/1094428102239425

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Belmont CA: Wadsworth group/Thomson learning.

Çelen, Ü., & Aybek, E. C. (2013). Öğrenci başarısının öğretmen yapımı bir testle klasik test kuramı ve madde tepki kuramı yöntemleriyle elde edilen puanlara göre karşılaştırılması. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *4*(2), 64-75. Retrieved from http://dergipark.gov.tr/epod/issue/5800/77213.

De Ayala, R. J., & Sava-Bolesta, M. (1999). Item parameter recovery for the nominal response model. *Applied Psychological Measurement*, *23*(1), 3-19. doi: 10.1177/01466219922031130

DeMars, C. E. (2002, April). *Recovery of graded response and partial credit parameters in MULTILOG and PARSCALE*. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

DeMars, C. E. (2003). Sample size and the recovery of nominal response model item parameters. *Applied Psychological Measurement*, *27*(4), 275-288. doi: 10.1177/0146621603027004003

Doğan, N., & Tezbaşaran, A. A. (2003). Klasik test kuramı ve örtük özellikler kuramının örneklemler bağlamında karşılaştırılması. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, *25*(25), 58-67. Retrieved from http://dergipark.gov.tr/download/article-file/87861.

Dolma, S. (2009). *Çok ihtimalli Rasch modeli ile derecelendirilmiş yanıt modelinin örtük özellikleri tahminleme performansı açısından simülasyon yöntemiyle karşılaştırılması* (Yayımlanmamış Doktora Tezi, İstanbul Üniversitesi Sosyal Bilimler Enstitüsü, İstanbul). Erişim adresi: https://tez.yok.gov.tr/UlusalTezMerkezi/.

Fotaris, P., Mastoras, T., Mavridis, I., & Manitsaris, A. (2010, September). Performance evaluation of the small sample dichotomous IRT analysis in assessment calibration. In *Computing in the Global Information Technology (ICCGI), 2010 Fifth International Multi-Conference on* (pp. 214-219). IEEE. doi: 10.1109/ICCGI.2010.19

Guyer, R., & Thompson, N. (2011). *Item response theory parameter recovery using Xcalibre 4.1*. Saint Paul, MN: Assessment Systems Corporation.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory principles and applications* (2. Ed.). USA: Kluwer-Nijhoff.

Hambleton, R. K., Swaminathan, H. & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.

Han, K. T. (2007). WinGen: Windows software that generates IRT parameters and item responses. *Applied Psychological Measurement, 31*(5), 457-459. doi: 10.1177/0146621607299271

Han, K. T., & Hambleton, R. K. (2007). *User's manual: WinGen* (Center for Educational Assessment Report No. 642). Amherst, MA: University of Massachusetts, School of Education.

Han, K. T. (2012). Fixing the c parameter in the three-parameter logistic model. *Practical Assessment, Research & Evaluation, 17*(1), 1-24. Retrieved from http://pareonline.net/getvn.asp?v=17&n=1.

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

İlhan, M. (2016). Açık uçlu sorularla yapılan ölçmelerde klasik test kuramı ve çok yüzeyli Rasch modeline göre hesaplanan yetenek kestirimlerinin karşılaştırılması. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 31*(2), 346-368. doi: 10.16986/HUJE.2016015182

Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982). Recovery of two-and three-parameter logistic item characteristic curves: A Monte Carlo study. *Applied Psychological Measurement, 6*(3), 249-260. doi: 10.1177/014662168200600301

Kieftenbeld, V., & Natesan, P. (2012). Recovery of graded response model parameters: A comparison of marginal maximum likelihood and Markov chain Monte Carlo estimation. *Applied Psychological Measurement, 36*(5), 399–419. doi: 10.1177/0146621612446170

Koğar, H. (2015). Madde tepki kuramına ait parametrelerin ve model uyumlarının karşılaştırılması: Bir Monte Carlo çalışması. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *6*(1), 142-157. doi: 10.21031/epod.02072

Köse, İ. A. (2010). *Madde tepki kuramına dayalı tek boyutlu ve çok boyutlu modellerin test uzunluğu ve örneklem büyüklüğü açısından karşılaştırılması* (Yayınlanmamış Doktora Tezi, Ankara Üniversitesi Eğitim Bilimleri Enstitüsü, Ankara). Retrieved from https://tez.yok.gov.tr/UlusalTezMerkezi/.

Lautenschlager, G. J., Meade, A. W., & Kim, S. H. (2006, April). *Cautions regarding sample characteristics when using the graded response model.* Paper presented at the 21st Annual Conference of the Society for Industrial and Organizational Psychology, Dallas, TX.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* USA: Information Age Publishing.

Martin, M. O., Mullis, I. V. S., & Hooper, M. (Eds.). (2016). Methods and procedures in TIMSS 2015. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College and International Association for the Evaluation of Educational Achievement (IEA).

Masters, G. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149–174. doi: 10.1007/BF02296272

Montgomery, M., & Skorupski, W. (2012, April). *Investigation of IRT parameter recovery and classification accuracy in mixed format.* Paper presented at the annual meeting of the Nation Council of Measurement in Education, British Columbia.

Muraki, E. (1992) A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159–176. doi: 10.1002/j.2333-8504.1992.tb01436.x

OECD. (2017). *PISA 2015 Technical Report*. Paris: PISA, OECD Publishing.

Preinerstorfer, D., & Formann, A. K. (2012). Parameter recovery and model selection in mixed Rasch models. *British Journal of Mathematical and Statistical Psychology*, *65*(2), 251-262. doi: 10.1111/j.2044-8317.2011.02020.x

Preston, K. S. J., & Reise, S. P. (2014). Estimating the nominal response model under nonnormal conditions. *Educational and Psychological Measurement*, *74*(3), 377-399. doi: 10.1177/0013164413507063

Reise, S. P., & Yu, J. (1990). Parameter recover in the graded response model using MULTILOG. *Journal of Educational Measurement, 27*(2), 133-144. doi: 10.1111/j.1745-3984.1990.tb00738.x

Roberts, J. S., & Laughlin, J. E. (1996). A unidimensional item response model for unfolding responses from a graded disagree-agree response scale. *Applied Psychological Measurement, 20*(3), 231-255. doi: 10.1177/014662169602000305

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, 34*(4, Pt. 2), 100.

Seong, T. J., Kim, S. H., & Cohen, A. S. (1997, March). *A comparison of procedures for ability estimation under the graded response model.* Paper presented at the annual meeting of the Nation Council of Measurement in Education, Chicago.

Sen, S. (2014). *Robustness of mixture IRT models to violations of latent normality* (Doctoral dissertation, University of Georgia, Athens). Retrieved from http://tez.yok.gov.tr/ UlusalTezMerkezi/.

Sen S., Cohen A.S., Kim SH. (2015) Robustness of Mixture IRT Models to Violations of Latent Normality. In: Millsap R., Bolt D., van der Ark L., Wang WC. (eds) Quantitative Psychology Research. Springer Proceedings in Mathematics & Statistics, vol 89. Springer, Cham. doi: 10.1007/978-3-319-07503-7_3

Stone, C. A. (1992). Recovery of marginal maximum likelihood estimates in the two- parameter logistic response model: An evaluation of MULTILOG. A*pplied Pscyhological Measurement, 16*(1), 1-16. doi: 10.1177/014662169201600101

Swaminathan, H. & Gifford, J. A. (1979, April). *Estimation of parameters in the three- parameter latent trait model*. Paper presented at the annual meeting of AERA-NCME, San Francisco.

Tate, R. (2000). Robustness of the school-level polytomous IRT model. *Educational and Psychological Measurement, 60*(1), 20-37. doi: 10.1177/00131640021970349

Thissen, D., Chen, W. H. & Bock, D. (2003). *MULTILOG 7.03*. Lincolnwood, IL: Scientific Software International.

Thissen, D., & Wainer, H. (1982). Some standard errors in item response theory. *Psychometrika*, *47*(4), 397-412. doi: 10.1007/BF02293705

Thompson, N. A., & Weiss, D. J. (2011). A framework for the development of computerized adaptive tests. *Practical Assessment, Research & Evaluation, 16*(1), 1-9. Retrieved from http://pareonline.net/getvn.asp?v=16&n=1.

Wang, W. C., & Chen, C. T. (2005). Item parameter recovery, standard error estimates, and fit statistics of the Winsteps program for the family of Rasch models. *Educational and Psychological Measurement, 65*(3), 376-404. doi: 10.1177/0013164404268673

Wollack, J. A., Bolt, D. M., Cohen, A. S., & Lee, Y. S. (2002). Recovery of item parameters in the nominal response model: A comparison of marginal maximum likelihood estimation and Markov chain Monte Carlo estimation. *Applied Psychological measurement, 26*(3), 339-352. doi: 10.1177/0146621602026003007

Wollack, J. A., & Cohen, A. S. (1998). Detection of answer copying with unknown item and trait parameters. *Applied Psychological Measurement, 22*(2), 144-152. doi: 10.1177/01466216980222004

Yavuz, G., & Hambleton, R. K. (2017). Comparative analyses of MIRT models and software (BMIRT and flexMIRT). *Educational and Psychological Measurement, 77*(2), 263-274. doi: 10.1177/0013164416661220

Farklı Yetenek Dağılımlarında Madde Güçlük Dağılımı, Test Uzunluğu ve Örneklem Büyüklüğünün İncelenmesi

Giriş

Madde tepki kuramının (MTK) karakteristik özellikleri sayesinde bireye uygun test geliştirme, madde yanlılığını belirleme, testleri eşitleme gibi durumlarda ilerleme sağlanmış, sınırlılıklar giderilmiştir (Hambleton ve Swaminathan, 1985). MTK'nın birçok avantajından dolayı PISA, TIMSS gibi uluslararası sınavlarda tercih edildiği görülmektedir. Ayrıca ulusal ve uluslararası birçok araştırmada sınavlardan elde edilen sonuçların MTK bağlamında değerlendirildiği de görülmektedir. Bireyler için oldukça önemli bir konu olan ve eğitimde kullanılan sınavlar farklı amaçlarla hazırlanmaktadır. Bu amaçlar arasında öğrencileri seçme ve yerleştirme, düzey belirleme, girdi özelliklerini belirleme, öğrencileri sıralama vb. yer alabilir. Sınavlar hazırlanış ve uygulanış amacına veya testi alan bireylerin özelliklerine ve /veya sayısına göre farklı psikometrik özelliklere de sahip olacaktır. Örneğin bir testi alan birey sayısının fazla fakat test sonucu ile karar verilecek birey sayısı az ise hazırlanan testin zor olması beklenen bir durumdur. Ancak seçme ve yerleştirme amacından çok bireylerin var olan bilgilerinin tespiti için hazırlanan bir sınavın ise seçme ve yerleştirme sınavına göre daha kolay olması hatta mümkünse çoğunluğunun orta güçlükte maddelerden oluşması daha istendik bir durumdur. Burada asıl olan testlerde ölçme ve değerlendirme açısından sağlanması gereken geçerlik ve güvenirliğin bu durumdan nasıl etkileneceğinin belirlenmesidir. Ayrıca testi alan bireylerin yetenek dağılımlarının farklılaşmasının da geçerlik ve güvenirliğe olan etkisinin belirlenmesi de önemlidir.

Bu çalışmada ulusal bir sınavdan elde edilen parametrelere dayanarak birey dağılımının sağa ve sola çarpık olması durumunda, farklı b parametresi dağılımlarının, test uzunluğunun ve örneklem büyüklüğünün birey parametresi kestiriminde ölçme kesinliğine etkisi incelenmiştir. Literatürde

birey dağılımı türü, örneklem büyüklüğü ve test uzunluğu koşullarının ölçme kesinliğine etkisinin incelendiği sıklıkla görülmektedir. Ancak farklı birey dağılımları, test uzunlukları ve örneklem büyüklüklerinde farklı *b* parametresi dağılımlarının ölçme kesinliğine etkisinin incelendiği çalışmalara literatürde rastlanmamıştır. Burada farklı madde güçlüğü dağılımlarına dayalı olarak türetilen dört farklı testin işe koşulması çalışmanın ayrıca önemini oluşturmaktadır.

1. Sağa çarpık yetenek dağılımında, farklı test uzunlukları, örneklem büyüklükleri ve madde güçlük dağılımlarının yetenek parametresi kestiriminin ölçme kesinliğine etkisi nedir?
2. Sola çarpık yetenek dağılımında, farklı test uzunlukları, örneklem büyüklükleri ve madde güçlük dağılımlarının yetenek parametresi kestiriminin ölçme kesinliğine etkisi nedir?

Yöntem

Araştırma kapsamında kullanılan koşulların oluşturulması amacıyla veriler üretildiğinden bu çalışma simülasyon çalışmasıdır. Araştırmada öncelikle birey parametreleri elde edilmiştir. Bu amaçla, liselere geçişte uygulanan ulusal öğrenci seçme sınavının 20 maddelik matematik alt testinden elde edilen veriler kullanılmıştır. Araştırmada 500, 1000, 2500, 5000 ve 10000 olmak üzere toplam beş örneklem büyüklüğü belirlenmiştir. Simülasyon çalışması için ilk aşamada gerçek birey parametreleri elde edilmiştir. Sağa çarpık birey parametrelerinin elde edilmesinde her bir örneklem büyüklüğü için gerçek veriden random gruplar seçilmiştir. Sola çarpık birey parametrelerinin elde edilmesinde ise verinin tamamından kasıtlı örnekleme yoluyla çarpıklık ≈-1,00 olacak şekilde her örneklem büyüklüğünde veri setleri seçilmiştir. Simülasyonun 2. aşamasında ise madde parametreleri türetilmiştir. Bu aşamada farklı b parametresi dağılımına sahip (normal dağılım, tekdüze dağılım, sola çarpık ve sağa çarpık dağılım) hem 20 maddelik hem de 30 maddelik testler oluşturulmuştur. Madde parametrelerinin üretilmesinde a parametre değeri min= 0,5 maks=2 olarak, c parametre değeri min= 0 maks=0,05 olarak belirlenmiştir. Sola çarpık b parametresi dağılımı için α=8; β=2; sağa çarpık b parametresi dağılımı için α=2; β=8; tekdüze b parametre dağılımı için min=-3; maks=+3; normal b parametresi dağılımı için ort=0; Ss= 1 değerleri kullanılarak araştırma kapsamında kullanılacak dört ayrı madde güçlüğü dağılımı oluşturulmuştur.

Araştırma kapsamına alınan 80 koşul (2 birey dağılımı x 5 örneklem büyüklüğü x 4 b parametresi dağılımı x 2 test uzunluğu) Wingen 3 programı (Han, 2007) yardımıyla oluşturulmuştur. MTK'de parametre iyileştirme çalışmalarında genel olarak ölçme kesinliği hesaplaması yapılmaktadır. Ölçme kesinliğini incelemek amacıyla "hata kareleri ortalamasını karekökü" (Root Mean Squared Error (RMSE)) ve "ortalama mutlak farklılık" (Absolute Average Deviation (AAD)) değerleri hesaplanmıştır.

1. Alt probleme ilişkin bulgular: Sağa çarpık birey dağılımında ele alınan tüm örneklem büyüklüklerinde ölçme kesinliği en yüksek; b parametresi dağılımı normal ve test uzunluğu 30 madde olduğunda, en düşük ise b parametresi sağa çarpık ve test uzunluğu 20 madde olduğunda elde edilmiştir. Ayrıca ölçme kesinliğinin normal b dağılımdan sonra en yüksek tekdüze b dağılımında olduğu gözlemlenmiştir. Araştırmanın sonuçları test uzunluğu açısından incelendiğinde ise, normal, tekdüze ve sağa çarpık b dağılımlarında genel olarak 20 maddelik teste ilişkin ölçme kesinliğinin 30 maddelik teste göre daha düşük olduğu belirlenmiştir. Bu b dağılımlarının aksine sola çarpık b dağılımda ise 20 maddelik testin ölçme kesinliğinin 30 maddelik teste göre daha yüksek olduğu görülmüştür. Sonuç olarak test uzunluğu arttıkça ölçme kesinliğinin de arttığı belirlenmiştir. Son olarak örneklem büyüklüğünün birey parametresinin kestiriminde ölçme kesinliğine önemli bir etkisinin olmadığı gözlemlenmiştir.
2. Alt probleme ilişkin bulgular: Sola çarpık birey dağılımında ele alınan farklı test uzunluklarında ve örneklem büyüklüklerinde b parametresi dağılımı normal olduğunda ölçme kesinliğinin en yüksek düzeyde olduğu ve bunu tekdüze dağılımın takip ettiği

söylenebilir. Ayrıca en düşük ölçme kesinliğinin de tüm test uzunluğu ve örneklem büyüklüklerinde en düşük sola çarpık b dağılımında olduğu görülmüştür. Son olarak sola çarpık birey dağılımı için örneklem büyüklüğünün ve test uzunluğunun birey parametrelerinin kestirim üzerinde önemli bir etkisi olmadığı gözlemlenmiştir.

Sonuç ve Tartışma

Araştırmadan elde edilen sonuçlarda, hem sağa hem de sola çarpık birey dağılımda farklı b dağılımına sahip her bir test için örneklem büyüklüğü arttıkça ölçme kesinliği için hesaplanan RMSE ve, AAD değerlerinde çok fazla değişim olmadığı görülmüştür. Sağa çarpık birey dağılımı için tüm örneklem büyüklüklerinde test uzunluğunun etkisi incelendiğinde ise test uzunluğu arttığında RMSE ve AAD değerlerinin genel olarak azaldığı gözlemlenmiştir. Ancak sola çarpık birey dağılımı için test uzunluğundaki değişimin ölçme kesinliğini önemli derecede etkilemediği görülmüştür. Ayrıca sağa ve sola çarpık birey dağılımlarında, tüm örneklem büyüklüğü ve test uzunlukları için; en yüksek ölçme kesinliği b parametresi dağılımı normal olduğunda elde edilmiştir. Normal b dağılımını ise b parametresinin tekdüze dağıldığı koşul izlemiştir. Son olarak sağa çarpık birey dağılımı için RMSE ve AAD değerlerinin en yüksek sağa çarpık *b* dağılımında olduğu, sola çarpık birey dağılımda ise en yüksek sola çarpık b dağılımında olduğu gözlemlenmiştir.

Araştırmanın sonuçları doğrultusunda test geliştiricilere sola çarpık *b* parametre dağılımı yani maddelerin çoğunluğunun zor olması ya da sağa çarpık *b* parametre dağılımı yani maddelerin çoğunluğunun kolay olması önerilmez. Çünkü bu tip *b* parametresi dağılımlarında ölçme kesinliği normal ve tekdüze *b* parametresi dağılımına kıyasla daha düşük elde edilebilmektedir. Başka araştırmalarda örneklem büyüklüğü ve test uzunluğu yerine kestirim yöntemi, model, çoklu puanlanan maddeler için kategori sayısı, tekrar sayısı, kestirim programı vb. gibi koşulların ölçme kesinliğine etkisi incelenebilir. Ayrıca yetenek parametreleri normal ve tekdüze dağılıma sahip olduğunda, farklı *b* parametresi dağılımlarının ölçme kesinliğine etkisi de araştırılabilir.