International Journal of Evaluation and Research in Education (IJERE)
Vol. 8, No. 3, September 2019, pp. 425~433
ISSN: 2252-8822, DOI: 10.11591/ijere.v8i3.19807

Parameter estimation bias of dichotomous logistic item response theory models using different variables

Alper Köse[1], C. Deha Doğan[2]
[1]Department of Measurement and Evaluation, Abant İzzet Baysal University, Turkey
[2] Department of Measurement and Evaluation, Ankara University, Turkey

Article Info

Article history:

Received Jun 14, 2019
Revised Aug 15, 2019
Accepted Aug 27, 2019

Keywords:

Ability distribution
Item response theory
Parameter recovery
Simulation

ABSTRACT

The aim of this study was to examine the precision of item parameter estimation in different sample sizes and test lengths under three parameter logistic model (3PL) item response theory (IRT) model, where the trait measured by a test was not normally distributed or had a skewed distribution.In the study, number of categories (1-0), and item response model were identified as fixed conditions, and sample size, test length variables, and the ability distributions were selected as manipulated conditions. This is a simulation study. So data simulation and data analysis were done via packages in the R programming language. Results of the study showed that item parameter estimations performed under normal distribution were much stronger and bias-free compared to non-normal distribution. Moreover, the sample size had some limited positive effect on parameter estimation. However, the test length had no effect parameter estimation. As a result the importance of normality assumptions for IRT models were highlighted and findings were discussed based on relevant literature.

Corresponding Author:

Alper Köse,
Department of Measurement and Evaluation,
Faculty of Education, Abant İzzzet Baysal University,
Bolu Abant İzzet Baysal Üniversitesi Eğitim Fakültesi 14280, Bolu, Turkey.
Email: i.alper.kose@gmail.com

1. INTRODUCTION

Researchers working in the fields of education and psychology have developed different measurement theories and models to explain the latent trait underlying individuals' response to an item. One of the fundamental and still widely adopted examples is the Classical Test Theory (CTT), which aims to explain the score obtained from a test based on true and error scores. CTT contains easy-to-meet but weak assumptions. Despite CTT having served the field for many years, item and ability parameters being dependent on the group, difficulties in comparing individuals, and the presence of lower thresholds for reliability estimations [1] have led researchers to seek different measurement theories. Furthermore, discussions about true and observed scores used in CTT not having the same meaning as ability scores, and the literature concerning the necessity of ability scores to be independent from the test and the test items, contradicting the nature of CTT [2] have formed the foundations of a new theory.

After CTT, item response theory (IRT) is one of the most important theories, which have a place in the history of psychometry. IRT is a powerful test theory that explains the latent traits of probabilistic models underlying an individual's response to an item through much stronger assumptions compared to CTT [3, 4]. The three fundamental assumptions of IRT: unidimensionality, local independence, and monotonicity.

In the literature on IRT, the assumption that there is predominantly a single latent feature underlying the response of testers to a group of items in the test is referred to as unidimensionality. Undoubtedly, other factors, such as cognitive function, excitement, and stress also affect the test performance of individuals.

Therefore, the *predominant* factor responsible for the test performance of individuals is considered to be the trait measured by the test [1, 5]. Mathematical models explaining the performance of individuals in test items using multiple latent traits are known as IRT models in the related literature.

Local independence is the assumption that the probability of individuals responding to a test item in a certain way is independent of the probability of their response to other items in the same test, given that the latent trait measured by the test is kept constant. Although the literature contains specific mathematical calculations to meet this assumption, local independence is usually considered parallel to the unidimensionality assumption. This is because if the probability of individuals responding to each test item is independent of each other, then the only factor that explains the response probability is the latent trait measured by a single-factor test [5].

The last basic assumption of IRT is monotonicity, which is intertwined with the item characteristic curve. Unlike CTT, in IRT, the individuals' ability and their probability to correctly respond to the items in the test are curvilinear, and this curvilinear relationship is displayed using the item characteristic curve (ICC). Researchers defining IRT as the non-linear regression of item performance on the trait measured by the test, stated that IRT assumed that as the ability of an individual increased, the probability of his/her correctly responding to the test items also monotonically increased [6].

2. UNIDIMENSIONAL ITEM RESPONSE THEORY MODELS

In education and psychology fields, there are mathematical equations explaining the behavior and responses of individuals by associating independent variables and dependent variables based on the model concept. In IRT, models can be classified according to the number of parameters included, dimensionality, and scoring system [7] as well as depending on the type of mathematical function; e.g., normal ogive and logistic.

The two-parameter normal ogive model, developed by Lord, is the first IRT model to explain the probability of an individual correctly responding to an item. This model explains the relationship between the probability of correctly responding to an item and the latent trait (Θ) using the areas under the normal distribution curve Figure 1is given that: the regression of the latent trait (Θ) on the individual's function in the i*th* item (Y_i') is linear; the conditional distribution of Y_i's are *normal* for each Θ value; and scatter of Y_i's included in regression is homeostatic. The probability of an individual correctly responding to an item is explained using the following equation: $\text{Pi}(\theta) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-x^2/x}.dx$

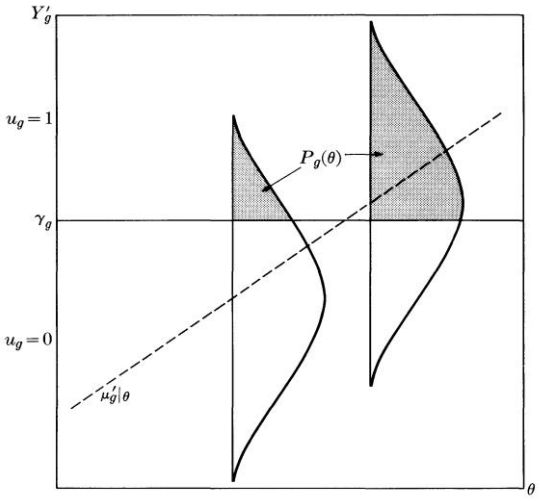

Figure 1. Normal ogive modelling

In this equation, a_i corresponds to $(\theta - b_i)$. In this first model of IRT, the calculation of $P(\theta)$ value requires complex integral calculations [8], which cannot even be undertaken using computer technologies. Therefore, researchers have sought new functions that eliminate the need for such complex integral calculations. As a result of these efforts, logistics models that emerged from Birnbaum's work in 1957

entered into the agenda of IRT. According to this model, the probability of an individual giving a correct response to an item can be expressed as:

$$Pi\ (\theta) = \frac{e^{Da(\theta-b)}}{1+e^{Da(\theta-b)}} \quad (1)$$

When the normal ogive and logistic function equations are examined, it is observed that the probability of an individual's response to an item is estimated based on the ability and item parameters. In addition, these equations provide researchers with important information concerning the item parameters included in the test and the abilities of individuals taking the test. This information allows work to be undertaken in relation to test development and the revision of a developed test. In order for these attempts to produce accurate results, the parameters included in the model must be estimated with a minimum error. In the estimation of model parameters, various approaches are utilized, such as maximum likelihood (ML), marginal maximum likelihood (MML), joint maximum likelihood (JML), full information maximum likelihood (FIML), and maximum a posteriori (MAP). The basic assumption of all these approaches is that the trait measured by a test (Ɵ) is normally distributed [9-11]. However, since this normality assumption is very difficult to be met, it is often accepted as normal [12, 13]. It has been observed that in cases where the trait measured by the test moves away from a normal distribution, and especially in the presence of a skewed distribution, item and ability parameter estimations are seriously affected and biased estimations may occur [14-17]. In addition, it is reported that particularly in nominal response models, 40% of the variability in item parameter estimation was explained by the latent trait distribution [6, 18, 19].

An examination of the related literature shows that although the factors affecting item parameter estimation; e.g., number of items, sample size, correlation between factors have been frequently investigated, the effects of latent trait distribution have not been adequately addressed. It is revealed that in data sets with a skewed latent trait distribution, as the number of items and sample size increased, the parameter estimation bias was reduced [20]. Consistent with this finding was the study by Kirisci Hsu and Yu who determined that the estimation bias in data sets with a skewed latent trait distribution could be largely compensated in cases of 40 items and 1,000 samples and above [10].

Throughout measurement history, various theories have been developed to explain the latent trait underlying the response of individuals to the items included in a measurement instrument. The aim of these theories is to perform an accurate and bias-free estimation of the parameters related to the items in the measurement tool and individuals responding to these items. In order to achieve these goals, such theories must meet some of their basic assumptions. Violations of these assumptions may cause significant problems in the estimation of the parameters mentioned above. The assumption of normal distribution, which is one of the basic assumptions of the substance response theory, is often one of the most difficult to be met. Thus, determining the accuracy (precision) and unbiasedness of estimation in cases where this assumption is met and is not met will make a significant contribution to the literature. In light of relevant literature and theoretical framework, the aim of this study was to examine the precision of item parameter estimation in different sample sizes and test lengths in cases where the trait measured by a test was not normally distributed or had a skewed distribution.

3. RESEARCH METHOD

This study aimed to examine the precision of item parameter estimation in different sample sizes and test lengths under non-normally distributed ability conditions. This is a simulation study. Simulation studies are computer experiments that involve creating data by pseudo-random sampling from known probability distributions. They are an invaluable tool for statistical research, particularly for the evaluation of new methods and for the comparison of alternative methods [21]. On the other hand since the current study was based on a theory test under various conditions, it would also be defined as a pure or fundamental research.

The data used in the research were produced using a simulation technique. The data were scored on a dichotomous basis (1-0) and prepared within the framework of the 3-P logistic model under IRT using the R programming language. When producing the data, it was ensured that the ability distribution was skewed. The results of the research were obtained by analyzing 100 replications of each condition.

In the study, number of categories (1-0), and item response model were identified as fixed conditions, and sample size, test length variables and the ability distributions, were selected as manipulated conditions.

For ability distribution, one of the independent variables selected in this study, the data were produced according to both normal and skewed distributions. In the normal distribution, ability distribution

was arranged to be mean = 0 and sd = 1. For the skewed distribution, the data were obtained in accordance with mean = 7 and sd = 2 using the "fGarch" package of R software.

One of the conditions affecting item parameter estimation in IRT is the sample size. It is suggested that estimation bias in skewed ability distributions could be compensated under the conditions of 40 items and 1,000 samples and above. For this reason, the sample sizes in this study were selected as 250, 500 and 1,000 [22].

In order to examine the effect of the test length variable under skewed ability distributions, 20-, 40- and 60-item test lengths were tested. As mentioned above, in the literature, it is stated that item parameter estimation errors in skewed distributions can be compensated when the test length is 40 items or above. Therefore, the test lengths of under and above 40 items were tested in this study.

Data Analysis: Bias and RMSE (Root Mean Square Error) calculations were used to evaluate the accuracy or precision of the estimated item parameters.

$$Bias = \frac{\sum_{i=1}^{K}(\hat{X}_i - X_i)}{K}$$
$$RMSE = \sqrt{\frac{\sum_{i=1}^{K}(\hat{X}_i - X_i)^2}{K}} \quad (2)$$

Where K is the test length, $\hat{X}_i$ refers to the estimated item parameter, and X_i represents the actual item parameter.

RMSE values being 0.05 or below indicates a good fit, while 0.10 or above indicates a poor fit of the model. Preferably, the bias value should be close to 0 [23-25].

4. RESULTS AND DISCUSSION

The findings obtained within the framework of the research problem are summarized in this section. The bias and RMSE values of b parameter estimation performed using the 1-PL model under different test length and sample size conditions and normal/non-normal distribution are reported in Table 1, Figure 2 and Figure 3.

Table 1. The bias and RMSE values of the b parameter under the 1-PL model

1 PL NORMAL DISTRIBUTION					1 PL NON-NORMAL DISTRIBUTION				
Sample Size	Test Length	Parameter	Bias	RMSE	Sample Size	Test Length	Parameter	Bias	RMSE
250	20	b	-0.01	0.162	250	20	b	-0.014	0.650
250	40	b	0.001	0.162	250	40	b	-0.003	0.624
250	60	b	-0.001	0.164	250	60	b	-0.006	0.591
500	20	b	-0.003	0.115	500	20	b	0.012	0.600
500	40	b	-0.002	0.117	500	40	b	0.012	0.640
500	60	b	-0.002	0.118	500	60	b	0.001	0.627
1000	20	b	0.001	0.008	1000	20	b	0.013	0.657
1000	40	b	-0.001	0.008	1000	40	b	-0.012	0.643
1000	60	b	0.001	0.009	1000	60	b	0.004	0.643

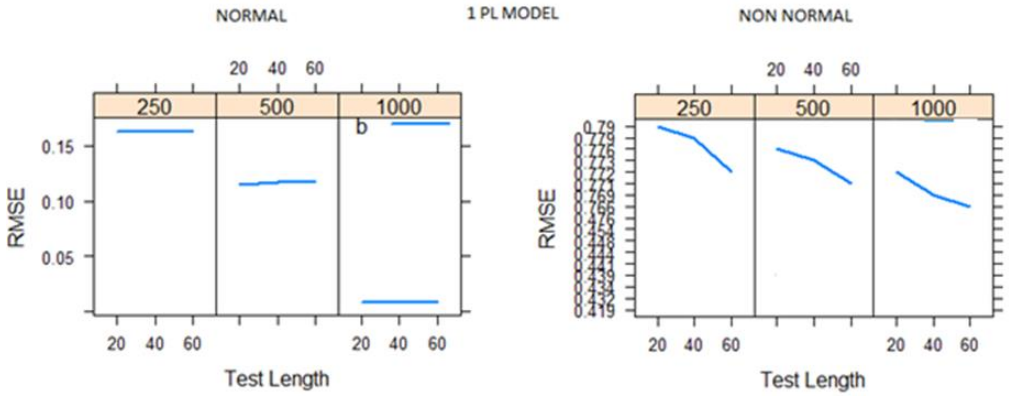

Figure 2. The RMSE values of the b parameter under the 1-PL model

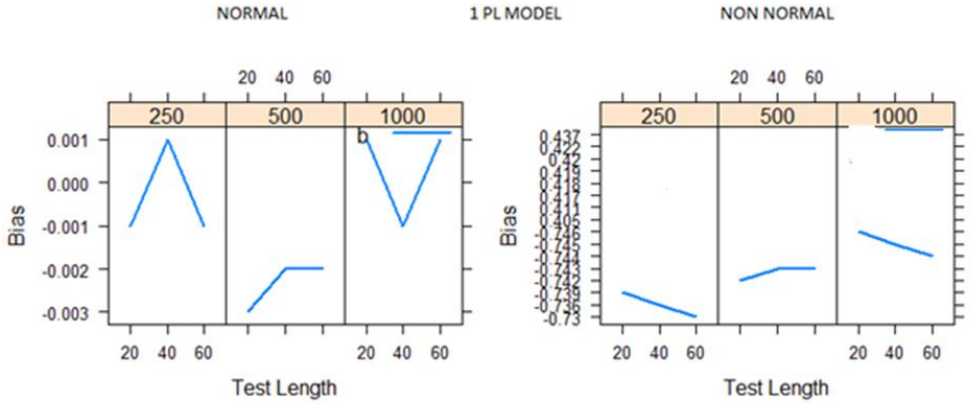

Figure 3. The bias values of the b parameters under the 1-PL model

The results presented in Table 1, Figures 2 and Figure 3 reveal that the RMSE and bias values estimated under normal distribution were lower and more acceptable than those obtained under non-normal distribution. As the sample size increased under normal distribution, the RMSE values were lower, and the bias values were not clear. When the length of the test was examined under both distributions, although no clear effect could be observed in normal distribution under the 1- PL model, better fit values were achieved as the number of items increased in non-normal distribution. Furthermore, the increase in both sample size and test length was found to have a positive effect on model fit values.

The bias and RMSE values of b and parameters estimation obtained using the 2-PL model based on different test lengths and sample sizes under normal and non-normal distributions are presented in Table 2, Figure 4 and Figure 5.

Table 2. The bias and RMSE values of the a and b parameters under the 2-PL model

2PL NON-NORMAL						2PL NORMAL					
Sample Size	Test Length	Bias a	Bias b	RMSE a	RMSE b	Sample Size	Test Length	Bias a	Bias b	RMSE a	RMSE b
	20	-0.74	0.418	0.790	0.454		20	0.023	-0.006	0.256	0.201
250	40	-0.74	0.411	0.779	0.448	250	40	0.025	-0.007	0.239	0.204
	60	-0.73	0.437	0.772	0.476		60	0.023	0.005	0.238	0.196
	20	-0.74	0.417	0.776	0.439		20	0.001	0.007	0.181	0.14
500	40	-0.74	0.419	0.733	0.441	500	40	0.014	0.000	0.165	0.135
	60	-0.74	0.422	0.771	0.444		60	0.007	0.006	0.163	0.34
	20	-0.75	0.405	0.772	0.419		20	0.012	-0.006	0.123	0.097
1000	40	-0.75	0.418	0.769	0.432	1000	40	0.004	-0.005	0.118	0.09
	60	-0.74	0.420	0.766	0.434		60	0.006	0.001	0.116	0.095

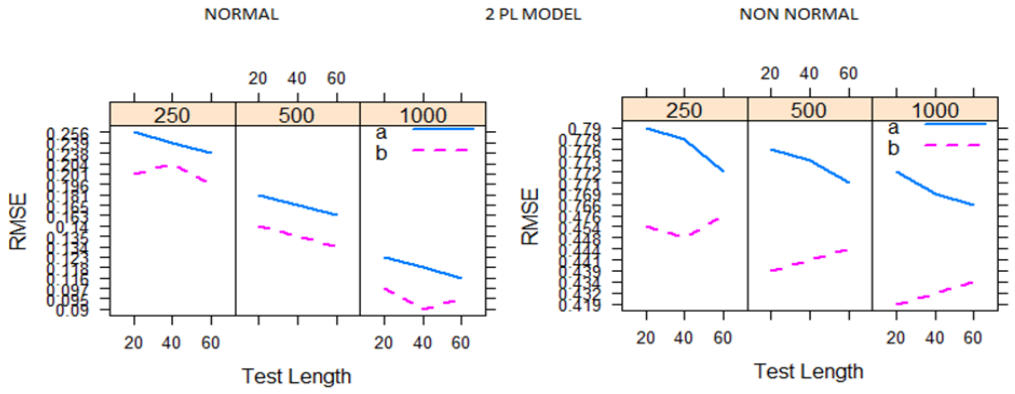

Figure 4. The RMSE values of the a and b parameters under the 2-PL model

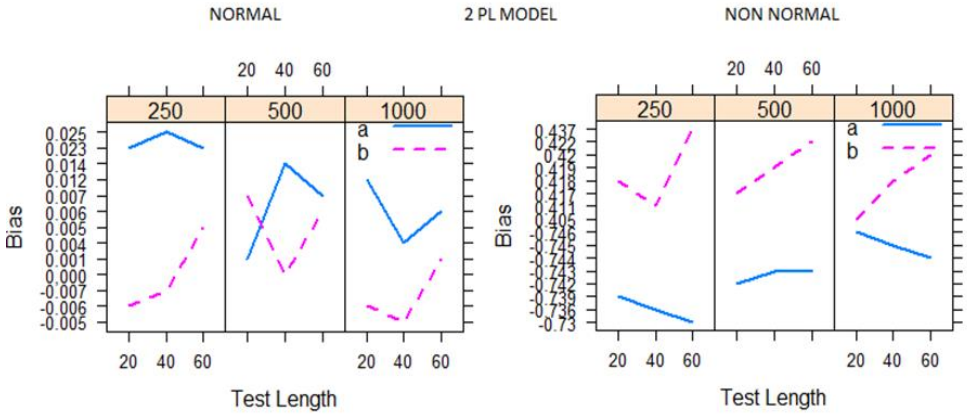

Figure 5. The bias values of the a and b parameters under the 2-PL model

According to the results given in Table 2 and Figure 4 and Figure 5, the RMSE and bias values of the a and b parameters estimated under normal distribution were more acceptable compared to those estimated under non-normal distribution. When the sample size variable was examined, it was observed that the RMSE and bias values were closer to 0, but the number of items did not have an effect that would result in consensus. Furthermore, the test length variable did not have a significant effect on model fit values.

The bias and RMSE values of b, a and c parameters estimation obtained using the 3-PL model based on different test lengths and sample sizes under normal and non-normal distributions are presented in Table 3, Table 4, Figure 6 and Figure 7.

Table 3. The bias and RMSE values of the a and b parameters of the non-normally distributed data under the 3-PL model

Sample Size	Test Length	Bias a	Bias b	Bias c	RMSE a	RMSE b	RMSE c
	20	0.451	-6.423	0.208	6.659	12.068	1.212
250	40	-0.141	-7.096	0.210	7.739	13.795	1.196
	60	-0.15	-7.491	0.217	8.135	14.893	1.195
	20	0.215	-4.761	0.215	4.436	9.054	1.186
500	40	-0.427	-4.542	0.204	4.667	9.369	1.213
	60	-0.53	-5.068	0.209	5.067	10.464	1.201
	20	-0.457	-3.073	0.193	2.961	6.445	1.216
1000	40	-0.838	-3.291	0.192	3.288	7.174	1.218
	60	-0.924	-3.220	0.189	3.252	7.251	1.221

Table 4. The bias and RMSE values of the a and b parameters of non-normally distributed data under the 3-PL model

Sample Size	Test Length	Bias a	Bias b	Bias c	RMSE a	RMSE b	RMSE c
	20	0.284	-0.278	0.052	0.683	2.069	1.344
250	40	0.284	-0.23	0.052	0.598	1.996	1.335
	60	0.303	-0.27	0.057	0.647	2.015	1.341
	20	0.148	-0.138	0.036	0.342	1.872	1.349
500	40	0.145	-0.162	0.040	0.318	1.870	1.344
	60	0.153	-0.152	0.041	0.317	1.844	1.348
	20	0.084	-0.082	0.029	0.229	1.716	1.361
1000	40	0.088	-0.083	0.029	0.206	1.794	1.357
	60	0.100	-0.105	0.030	0.208	1.78	1.356

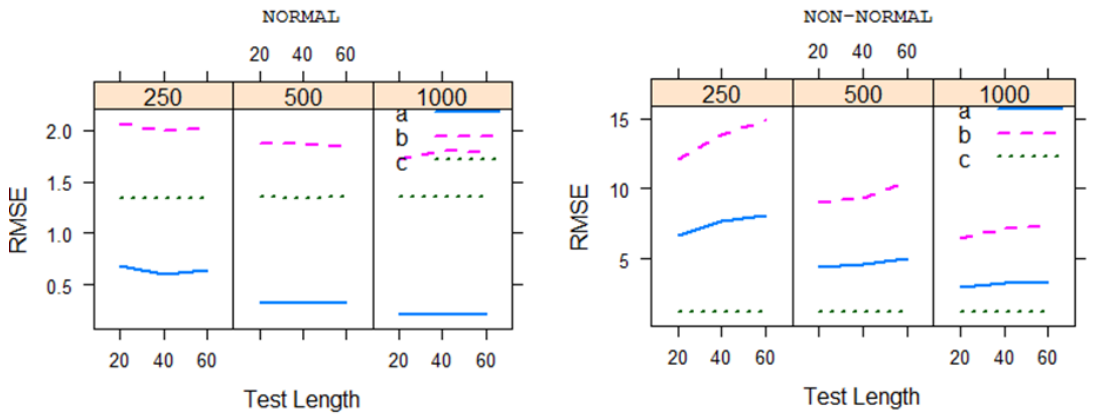

Figure 6. The RMSE values of the a, b and c parameters under the 3-PL model

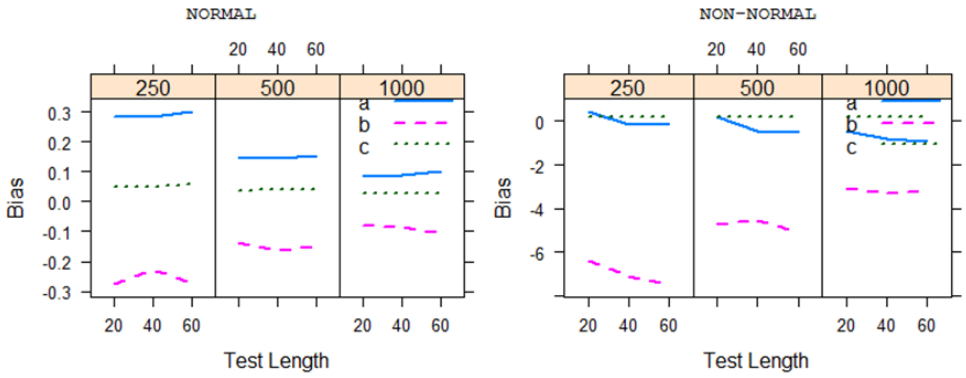

Figure 7. The bias values of the a, b and c parameters under the 3-PL model

Similar to the previous results, the RMSE and bias values of the a, b and c parameters estimated under normal distribution using the 3-PL model were much more acceptable than those estimated under non-normal distribution (Tables 3 and 4, Figure 6 and Figure7). Concerning the sample size variable, the RMSE and bias values were closer to 0, and the number of items had no significant on model fit values. As the sample size increased, the item discrimination parameter under both distributions produced more positive results.

5. DISCUSSION

In this study, the strength and bias of item parameter estimation were examined under two conditions by meeting and violating the normality distribution, one of the most important assumptions of IRT. The results obtained are discussed in relation to the literature.

The most significant result of this study was that item parameter estimations performed under normal distribution were much stronger and bias-free compared to non-normal distribution. There were very significant differences in the bias and RMSE values between the two distributions.

This finding is supported by other studies in the relevant literature [14-17]. To estimate the parameters of a test in IRT, various approached are adopted, such as ML, MML, JML, FIML, and MAP. The basic assumption of these approaches is that the trait measured by the test (Θ) is normally distributed [9-11]. This is the main reason why the RMSE and bias values were much more acceptable in the case of normal distribution.

Examining the variables of sample size and number of items, it could not be determined whether they had a consistently valid effect on each parameter in any of the conditions tested in this study. When the item difficulty parameter was positively affected by sample size in one model, the a parameter was more prominent in another estimation. It is reported that in data sets with a skewed latent trait distribution, as the number of items and sample size increased, parameter estimation bias was reduced [20]. In parallel to this,

Researchers determined that the estimation bias in data sets with a skewed latent trait distribution could be largely compensated in cases of 40 items and 1,000 samples or above [10, 22].

Although the findings of this study partially supported the relevant literature, no significant parallelism was found. This can be attributed to the differences in the data used. Although studies conducted with normally distributed data present similar results, the findings may differ in those based on data with a non-normal distribution. The conditions of the current study were different from other publications in the literature conducted under skewed distribution conditions. Thus, researchers using IRT should consider the structure of the ability distribution of the data when reporting the results of parameter estimation.

This study can be repeated in different conditions to further contribute to the literature. In addition, in recent years, IRT models that can be used in cases where the normality assumption is not met have been proposed in the literature. An example of this is the Ramsey Curve IRT model. In the future, studies can be conducted using this model to compare the results with the literature.

6. CONCLUSION

Results showed that item parameter estimations performed under normal distribution were much stronger and bias-free compared to non-normal distribution. There were very significant differences in the bias and RMSE values between the two distributions. Furhtermore the sample size had some limited positive effect on parameter estimation. However, the test length had no effect parameter estimation.

REFERENCES

[1] Hambleton, R. K., Swaminathan, H. and Rogers, H. *Fundamentals Of Item Response Teory*, Newbury Park CA: Sage,1991.
[2] Hambleton, R. K. ve Jones, R. W. "Comparison of classical test theory and item response theory and their applications to test development," *Educational Measurement*, vol. 12, pp. 38-47, 1993.
[3] Embretson, S.E. ve Reise, S.P. "Item Response Theory For Psychologists," *Lawrence Erlbaum Associate, Inc*,2000.
[4] Bobcock, B.G.E. "Estimating a Noncompensatory IRT Model Using a modified Metropolis algorithm," Unpublished Doctoral Dissertation.The University of Minesota, 2009.
[5] Hambleton, R.K. ve Swaminathan, H. *Item Response Theory. Principles And Applications*, Boston-USA: Kluwer-Nijhoff Publishing, 1989.
[6] Crocker, L. ve Algina, J., *Introduction to classical and modern test theory*, USA: Rinehart and Winston Inc,1986
[7] McDonald, R.P, "Linear Versus Models in Item Response Theory," *Applied Psychological Measurement*, vol. 6, pp. 379-396, 1982.
[8] Lord, F. M. ve Novick M. R. *Statistical theories of mental test scores*, New York: Addison- Wesley Publishing Company, 1968.
[9] Azevedo, C.L.N., Bolfarine, H. & Andrade, D.F. "Bayesian inference for a skew normal IRT model under the centered parameterization," *Computational Statictics and Data Analysis*, vol. 55, pp. 353-365, 2010.
[10] Finch, H. and Edwards, J.M. "Rasch model parameter estimmation in the presence of a nonnormal latent trait using a nonparametric Bayesian approach," *Educational and Psychological Measurement*, vol. 76(4), pp. 622-684, 2016
[11] Reise, S.P., Rodriguez, A., Spitzer, K.L. & Hays, R.D. "Alternative approaches to addressing non-normal distrib utions in the application of IRT models to personality measures," *Journal of Personality Assessment*, 2017.
[12] Micceri, T. "The unicorn, the normal curve and other improbable creatures," *Psychological Bulletin*, vol. 105(1), pp. 156-166, 1989.
[13] Samejima, F. "Departure from normal assumptions: a promise for future psychoemtrics with substantive mathematical modeling," *Psychometrica*, vol. 62(4), pp. 471-493, 1997.
[14] Seong, T. "Sensivity of marginal maximum likelihood estimation of item and ability parameters to the characteristic of the prior ability distributions," *Applied Psychological Measurement*, vol. 14(3), pp. 299-311, 1990.
[15] Woods, C.M. "Ramsay curve item response theory fort he three-parameter item response theory model," *Applied Psychological Measurement*, vol. 36, pp. 447-465, 2008.
[16] Woods,C.M. and Linn, N. "Item response theory with estimation of the latent density using Davidian curves," *Applied Psychological Measureent*, vol. 33, pp. 102-117, 2009.
[17] Woods, C.M. and Thissen, D "Item response theory with estimation of the latent population distribution using spline-based densities," *Psychometrica*, vol. 71, pp. 281-301,2006.
[18] DeMars, C.E. "Saple size and recovery of nominal response model item parameters," *Applied Psychological Measurement*, vol. 27(4), pp. 275-288, 2003.
[19] Wollack, L.A, Bolt, D.M., Cohen, A.S. and Lee, Y. "Recovery of item parameters in the nominal response model: A coparison of marginal maximum likelihood estimation and Markov Chain Monte Carlo Estimation," *Applied Psychological Measureent*, vol. 26(3), 339-352, 2002.
[20] Stone, C.A. "Recovery of marginal maximum lilkelihood estimates in the two parameter logistic model: An evaluation of MULTILOG," *Applied Psychological Measurement*, vol. 16, pp. 1-16, 1992.
[21] Moriis, T.P., White,I.R. and Crowther,M.J. "Using simulation studies to evaluate statistical methods," *Statistics in Medicine*, vol. 38(11), pp. 2074-2102, 2019.

[22] Kirisci, L., Hsu, T., and Yu, L. "Robustness of item parameter estimation programs to assumptions of unidimensionality and normality," *Applied Psychological Measurement*, 25(2), 146-162,2001
[23] Bulut, O. and Sünbül, Ö. "Monte carlo simulation studies in item response theory with the r programming language," *Journal of Measurement and Evaluation in Education and Psychology*, vol. 8(3), pp. 266-287, 2017.
[24] Hu, L. T., and Bentler, P. M. "Cutoff criteria for fit indices in covariance structure analysis: Conventional criteria versus new alternatives," *Structural Equation Modeling*, vol. 6, pp. 1-55, 1999.
[25] Millsap, R. E., and Everson, H. "Methodology review Statistical approaches for assessing measurement bias," *Applied Psychological Measurement*, vol. 17, pp. 297-334, 1993.