Celal Bayar University Journal of Science

Inferences from Bootstrap Method for Ability Parameters in 2-Parameter Logistic Model

Hülya Olmuş[1], Ezgi Nazman[2]*

[1]Gazi University, Faculty of Science, Statistics, Ankara, Turkey
[2]*Sivas Cumhuriyet University, Faculty of Science, Statistics and Computer Sciences, Sivas, Turkey
*ezgicabuk@cumhuriyet.edu.tr
*Orcid No: 0000-0003-0189-3923

Received: *20 September 2019*
Accepted: *14 September 2020*
DOI: *10.18466/cbayarfbe.622868*

Abstract

The ability parameter of persons/examinees estimates can be obtained using the Joint Maximum Likelihood (JML) estimation in Item Response Theory (IRT). However, JML estimates can be biased in some cases. Although the Bootstrap method has been considered for JML, existing studies remain far from satisfactory concerning the ability parameter estimation. This research evaluates the performances of JML and Bootstrap estimates of the ability parameter in terms of Standard Error Measurement (SEM) in the 2-Parameter Logistic (2-PL) model conducting a detailed Monte Carlo simulation study. According to the results, the average SEM estimates of the Bootstrap method are less than the average SEM estimates of JML in terms of the ability parameter.

Keywords: ability parameter, difficulty parameter, discrimination parameter, joint maximum likelihood estimation, two-parameter logistic model

1. Introduction

Measurement and evaluation methods have gained importance in many fields such as education, psychology, and medicine in past decades [1]. In general, researchers have considered measuring the ability parameter of persons/examinees (latent variable) such as intelligence, mathematical or scholastic abilities. Evaluating results of measurement techniques can be difficult in some fields where the latent variable has an important role. Classical Test Theory (CTT) is unable to assess the true measurement of the ability of examinees and the characteristics of items. The mathematical or verbal ability of students in education, consumer preferences in marketing, political attitude of voters in politics, etc. can be given as an example of the field where the ability parameter of examinees cannot be measured by CTT directly. Thus, Item Response Theory (IRT) has been widely used to estimate both items and examinee parameters in literature [2, 3].

IRT is a mathematical model that indicates the relation among examinee and item parameters and provides parameter estimates of both ability parameter and item parameters [4]. This model is based on the probability of giving the correct answer to a given specific item. This probability is the chance that the *i*-th examinee correctly answers to the *j*-th item and is denoted by $P(Y_{ij}=1|\theta_i)$

IRT models have three basic assumptions: unidimensionality, local independence, and monotonicity. The unidimensionality indicates that all items should measure only one examinee parameter. The local independence states that the probability of the correct response from the examinee is based solely on the ability of the examinee and each individual item, and not the interrelationship of multiple items. The monotonicity describes the functionality between an examinee's ability and performance on each item of the assessment [5].

IRT models are generalized linear models since they involve a transformation of the expected values with the help of a link function to depend on a linear formulation, and they are also mixed because one or more weights in the linear component are random variables [6]. Logit and Probit are widely used link functions in IRT. The logit link function is benefitted from standard logistic distribution and rather preferable function because of the computation easiness. The most widely used models in IRT are one-parameter logistic (1-PL), two-parameter logistic (2-PL) and three

Celal Bayar University Journal of Science
Volume 16, Issue 3, 2020, p 333-338
Doi: 10.18466/cbayarfbe.622868

parameter logistic (3-PL) models. 1-PL model is also known as Rasch model and contains only item discrimination parameter (a_j). 2-PL model, which is known as Birnbaum model, consist of both item discrimination and difficulty parameters [4]. In addition to the item discrimination and difficulty parameters, 3-PL model has a chance parameter (c_j) [7].

Item discrimination (a) and difficulty parameter (b) ranges from $-\infty$ to $+\infty$. However, it is assumed that, in practice, b value ranges from -3 to +3, when the examinee parameter (θ) has the standard normal distribution. The discrimination parameter indicates how well the item differentiates examinees. A higher discrimination parameter differentiates better among examinees. In the same way, a higher difficulty parameter indicates that the item is hard [1, 4]. The chance parameter (c) ranges from 0 to 1 and it is generally $c \leq 0.25$.

Liou and Yu [8] mentioned that the Bootstrap method can be used to determine the statistical accuracy of ability estimates in with given item parameters. Atanasov [9] studies on estimation of IRT parameters of the items with a small sample size using bootstrapping. Heene et al. [10] evaluated the performance of the Bootstrap for the Rasch model under the violations of non-intersecting item response functions. Wolfe and McGill [11] indicated that the Bootstrap critical values allow for greater statistical power in diagnosing item misfit caused by varying item slopes and lower asymptotes for the Rash model. Patton et al. [12] compared the performance of bootstrap standard error with the asymptotic standard error under 20 and 40 items for 500 and 2000 samples. Olmuş and Nazman [13] evaluated parameter estimations of 2-PL model using JML estimation. Liu and Yang [14] proposed a resampling-based method, namely bootstrap calibration, to reduce the impact of the carry over sampling error on the interval estimates of ability parameter. Liu et al [15] reviewed Monte Carlo methods in the literature in recent years. Chen et al. [16] used pseudo-population bootstrap to perform in terms of relative bias and coverage probability. However, there is still need to clear the performance of the Bootstrap method for higher item numbers. Therefore, we considered the 2-PL model in order to evaluate JML and the Bootstrap ability parameter estimation considering Standard Error Measurement (SEM) conducting a detailed Monte Carlo simulation study. The study was organized as follows: The model, item information function, test information function and standard error of measurement were explained in the second section. Parameter estimation of the model, JML estimation and the Bootstrap method were presented in the third section. Monte Carlo simulation study and obtained results were shown in the fourth and fifth sections, respectively.

2. Materials and Methods

2.1. Two-Parameter Logistic (2-PL) Model

The most widely used model in IRT is two-parameter logistic (2-PL) model. Let's consider a testing situation in which n examinees answer to k items. Let i=1, …, n and j=1, …, k be the random variables associated with the response of the i th examinee to the j-th item. These responses can be binary or discrete with a number of categories. Y_{ij} is the response for the i-th examinee to the j-th item and assumed to be identical for each item in the test. Here θ_i denotes the ability parameter for i-th examinee.

$P(Y_{ij}=1|\theta_i)$ denotes for the chance that the i-th examinee correctly answers the j-th item. Logit term of this probability and 2-PL model equation were given below respectively [3, 7].

$$\text{logit}(P(Y_{ij}=1))=ln\left(\frac{P(Y_{ij}=1)}{1-P(Y_{ij}=1)}\,\middle|\,\theta_i\right)=a_j\left(\theta_i-b_j\right) \quad \textbf{(2.1)}$$

$$P(Y_{ij}=1\,|\,\theta_i)=\frac{exp\left[a_j\left(\theta_i-b_j\right)\right]}{1+exp\left[a_j\left(\theta_i-b_j\right)\right]} \quad \textbf{(2.2)}$$

or

$$P(Y_{ij}=1\,|\,\theta_i)=\frac{1}{1+exp\left[-a_j\left(\theta_i-b_j\right)\right]} \quad \textbf{(2.3)}$$

where

a_j : the item discrimination parameter of j th item

b_j : the item difficulty parameter of j th item

2.1.1. Item Information Function (IIF)

An examinee's unknown ability do not depend upon the examinee's responses to the items. On the other hand, an examinee's unknown ability depends only on the parameter values of k items [17]. In 2-PL model, the general interest is mostly the estimated value of ability parameter for an examinee. The amount of information based on an item is able to be computed for any ability level. Item Information Function (IIF) for 2-PL model is shown as in Eq. (2.4):

$$I_i\left(\theta,b_j,a_j\right)=a_j^2\,P_i\left(\theta,b_j\right)Q_i\left(\theta,b_j\right) \quad \textbf{(2.4)}$$

where

$P_i\left(\theta,b_j\right)=\frac{1}{1+exp\left[-a_j\left(\theta_i-b_j\right)\right]}$ and $Q_i\left(\theta,b_j\right)=1-P_i\left(\theta,b_j\right)$.

The value of discrimination parameter is required to compute IIF [3, 15].

2.1.2. Test Information Function (TIF)

A study such as survey or test is a set of items. Thus, the test information gives the ability level is computed from the sum of the item informations at that level. The Test Information Function (TIF) for 2-PL model is defined as in Eq.(2.5):

$$I_i(\theta_i)=\sum_{j=1}^{k} I_{ij}(\theta_i,b_j,a_j) = \sum_{j=1}^{k} a_j^2 P(\theta,b_j,a_j)Q(\theta,b_j,a_j), \quad i=1,2,...,n \tag{2.5}$$

In general, TIF tends to be higher than that for IIF [18].

2.1.3. Standard Error of Measurement (SEM)

The variance of ability estimate in the 2-PL model can be estimated as the reciprocal value of TIF at the ability estimate. This standard error of ability parameter is given as in Eq.(2.6) [18].

$$SEM(\theta)=\sqrt{1/\sum_{j=1}^{k} a_j^2 P_j(\theta,b_j,a_j)Q(\theta,b_j,a_j)} \tag{2.6}$$

2.2. Parameter Estimation of 2-PL Model

Let $y_{i1},y_{i2},...,y_{ik}$ be the dichotomous response variables of the i th examinee to k items, $\boldsymbol{a}=(a_1,a_2,...,a_k)$ and $\boldsymbol{b}=(b_1,b_2,...,b_k)$ be the vectors of discrimination and difficulty parameters. When we assume that an examinee taking the test responses each item independently, the probability of observing a particular response matrix of the i-th examinee is given as in Eq.(2.7):

$$P(Y_{i1}=y_{i1},...,Y_{ik}=y_{ik}/\theta_i,\boldsymbol{a},\boldsymbol{b})=\prod_{j=1}^{k} P(Y_{ij}=y_{ij}/\theta_i,\boldsymbol{a},\boldsymbol{b}) \tag{2.7}$$

Then the likelihood function for all responses of examinees become as in Eq.(2.8):

$$L(\theta,a,b)=\prod_{i=1}^{n} P_j^{y_{ij}}(1-P_j)^{1-y_{ij}} \tag{2.8}$$

This function represents the likelihood of obtaining the observed data as a function of the model parameters. Therefore, it is logical to estimate these model parameters using those values that maximize this likelihood function.

2.2.1. Joint Maximum Likelihood Estimation (JML)

One of the important tasks when a test is examined is to estimate these parameter values because actual values of item parameters in a test are unknown. In IRT, estimation of both item and ability parameters is a crucial process. In the case of the 2-PL model, the log-likelihood for examinees is shown in Eq.(2.9):

$$ln\,L(\theta,a,b)=\prod_{i=1}^{n}\prod_{j=1}^{k}\left[y_{ij}ln(P_{ij})+(1-y_{ij})ln(1-P_{ij})\right] \tag{2.9}$$

Its partial derivations are taken with respect to each parameter and set them to zero. The obtained equations are not linear. Therefore, the Newton-Rapson method is used to obtain item and ability parameter estimations. JML method is used to estimate both item and ability parameters treating the parameters as fixed parameters. In the first stage, the item parameters are estimated assuming known examinee abilities. In the second stage, it is assumed that the item parameter values are known for the estimated examinee's ability parameters. Then, the process yields estimates for both item and ability parameters [18].

2.2.2 Bootstrap Method

The bootstrap resampling method allows researchers to quantify uncertainty by calculating standard errors and confidence intervals and performing significance tests. They require fewer assumptions than traditional methods and generally give more accurate answers [19]. In this study, the bootstrap method steps for 2- PL model as shown [12]:

Step 1: Estimate item and person parameters based on the original sample.
Step 2: Select the values of the item and person parameters randomly from the estimated values.
Step 3: Generate simulated data sets that fit the 2-PL model for each the bootstrap resampling.
Step 4: Compute the statistics of interest (the average estimates of the item and ability parameters) for each of the resamples.
Step 5: Compute averages of the statistics of interest across the bootstrap rasamples.
Step 6: Compare the value of the statistics of interest to the average bootstrap values.

3. Simulation Study

A Monte Carlo simulation study was conducted to compare estimated ability parameter with JML and Bootstrap method using MATLAB R2017b. Item numbers and sample sizes were determined as 60, 90, 180, and 150, 500, 1000, respectively. The ability parameters and the item difficulty parameter values were generated from N(0,1). Item discrimination parameters were randomly selected from the possible values of {0.2, 0.4, 0.6, 0.8, 1.0, 1.2, 1.4, 1.6}. First, the data set of 0's and 1's were generated using n ability examinees and k items each with two parameters. Using $n*k$ data matrix of simulated responses, the JML estimates were obtained for item and ability parameters

Celal Bayar University Journal of Science
Volume 16, Issue 3, 2020, p 333-338
Doi: 10.18466/cbayarfbe.622868

[20, 21]. In addition, the estimated values of ability estimation and SEM of ability parameters were obtained using the Bootstrap method when all item and ability parameters were estimated, SEM of the ability parameter estimates of JML were compared with SEM of the ability parameter estimates of the Bootstrap method. With this aim, we run 100 bootstrap resampling for each examinee. For each bootstrap resampling, item and ability parameter values were estimated using JML estimates again.

4. Results and Discussion

Estimation and standard error values (SEM) for the ability parameter of JML and the Bootstrap were given in Table1. The average SEM estimates for the ability and item parameters were estimated by using JML and the Bootstrap was given in Table1.

Table 1. Average estimations for ability parameter and SEM using JML and bootstrap methods.

n	k	a	b	θ	SEM(θ)	a_{boot}	b_{boot}	θ_{boot}	SEM(θ_{boot})
150	60	1.0348	0.0040	-0.0032	0.0227	1.2378	-0.0218	0.0038	0.0199
	90	0.9177	-0.1941	0.0098	0.0205	1.0519	-0.1677	0.0194	0.0184
	180	0.9252	0.0119	-0.0009	0.0144	1.0139	0.0180	-0.0083	0.0133
500	60	0.9408	-0.0093	-0.0018	0.0135	1.0644	-0.0166	-0.0118	0.0124
	90	0.9328	-0.0423	0.0004	0.0109	1.0196	-0.0340	0.0084	0.0103
	180	0.9289	0.0988	-0.0033	0.0078	0.9860	0.0941	-0.0139	0.0074
1000	60	0.9630	-0.1698	0.0043	0.0096	1.0892	-0.1634	0.0066	0.0086
	90	0.9786	-0.1141	0.0022	0.0076	1.0658	-0.1225	0.0030	0.0071
	180	0.9551	0.0470	-0.0042	0.0054	1.0119	0.0544	-0.0044	0.0052

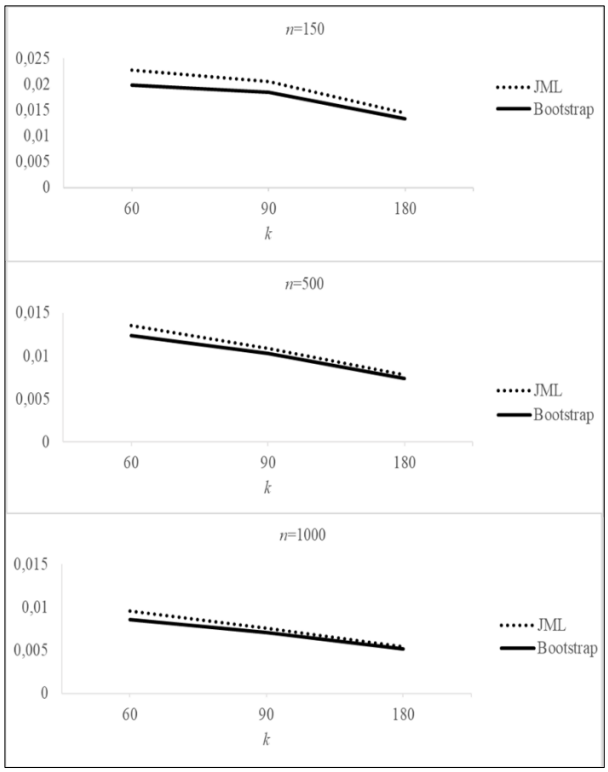

Figure 1. Average JML and Bootstrap estimates of ability parameter.

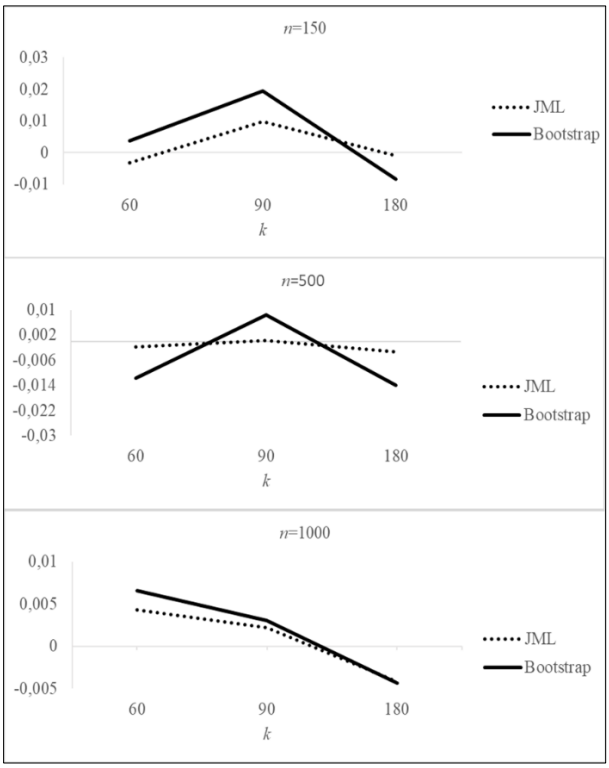

Figure 2. Average JML and bootstrap SEM estimates of ability parameter.

The Figure1 shows the variations of the discrimination parameter for various sample sizes and item numbers in this study.

3.1. Comments on Discrimination Parameter:

As shown in Figure2, when the number of items was low (k=60), for all sample sizes, it was seen that the average JML estimates of item discrimination

Celal Bayar University Journal of Science
Volume 16, Issue 3, 2020, p 333-338
Doi: 10.18466/cbayarfbe.622868

parameters were obtained less than the Bootstrap method. When the sample size is low (n=150), for all item numbers, it was obtained that the average discrimination parameter JML estimates were less than the Bootstrap method of item discrimination parameters. When the sample size increased for all item numbers, the average JML estimates differed from the average Bootstrap estimates in terms of item discrimination parameter. All in all, the Bootstrap method caused an increase in the average estimated value of the discrimination parameter. Figure 2 shows the variations of the difficulty parameter for various sample sizes and item numbers in this study. The major findings of this parameter are as follows:

3.2. Comments on Difficulty Parameter:

The average JML estimates of the item difficulty parameters tended to be so close with the average Bootstrap estimates of the item difficulty parameters when the sample size increased for item number 60, 90, and 180.

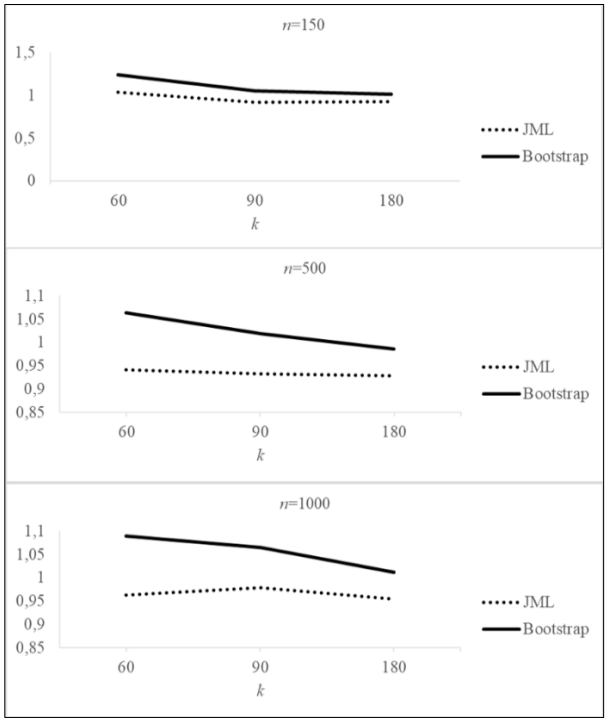

Figure 3. Average JML and Bootstrap estimates of discrimination parameter.

When the item number and sample size increased, an increase was observed in the average estimate values of the item difficulty parameter. The Figure3 shows the variations of the ability parameter for various sample sizes and item numbers.

3.3. Comments on Ability Parameter:

When the item number and sample size increased, the averages JML and Bootstrap estimates of ability parameters tended to be so close. Also, when the sample size and item number increased, the ability level of the examinee is on the decrease. When the sample size was and item number was low (n=150 and k=60), the ability level tended to increase. However, the average Bootstrap estimates were less than the average JML estimates in terms of ability parameters. Figure 4 shows the variations of the average SEM of the ability parameter for various sample sizes and item numbers in this study.

3.4. Comments on Standart Error (SEM) of Ability Parameter:

When the sample size increased for the low item number, the average SEM of JMLs was less than the average SEM of the Bootstrap estimates for the ability parameter. However, the average SEM of JMLs and the Bootstrap estimates were closer when the item number and sample sizes increase.

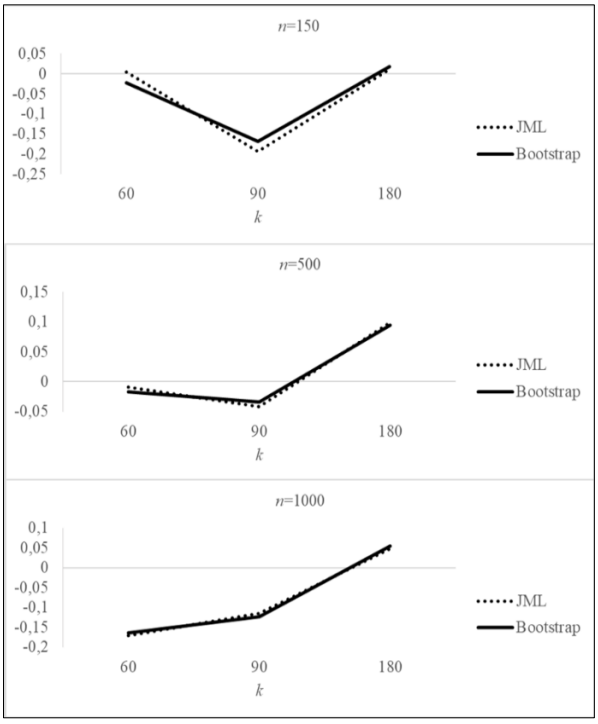

Figure 4. Average JML and Bootstrap estimates of difficulty parameter.

4. Conclusion

The major focus of this study was to compare the performances of Joint Maximum Likelihood and Bootstrap methods concerning the ability parameter and standard error measurement of ability parameter in two-parameter logistic model. It is well known that Joint

Maximum Likelihood estimates can be biased in some cases. According to the study, the Bootstrap is one of the approaches that may reduce the bias of Joint Maximum Likelihood estimates. It is seen that bootstrap method cause decrease on bias of ability parameter estimates in this study.

In general, item number has large impact on the accuracy of ability estimation than sample size. Therefore, Joint Maximum Likelihood and Bootstrap estimates give the same results when item number increase. The results show that in general for two-parameter logistic model, as item number increases, the accuracy of ability estimate measured by standard error measurement increases. However, it is result that bootstrap estimates causes an increase in estimation variability which can be shown in standard error measurement of ability parameter. It is seen that bootstrap method cause decrease on bias of estimated ability parameter.

Acknowledgement

Authors would like to thank the editor and two anonymous referees who kindly reviewed the earlier version of this manuscript and provided valuable suggestions and comments.

Author's Contributions

Hülya Olmuş: Drafted and wrote the manuscript, performed the experiment and result analysis.
Ezgi Nazman: Drafted and wrote the manuscript, performed the experiment and result analysis.

Ethics

There are no ethical issues after the publication of this manuscript.

References

1. Rasch, G. Probabilistic Models for Some Intelligence and Attainment Tests; Chicago: MESA; 1960.

2. Hambleton, RK, Jones, RW. 1993. Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*; 12(3): 38-47.

3. Baker, FB. The basis of item response theory. ERIC. 2001.

4. Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability, In Lord, FM, Novick, MR (Eds.), Statistical Theories of Mental Test Scores; 1968.

5. Paolino, JP. Penalized joint maximum likelihood estimation applied to two parameters logistic item. Columbia University Graduate School of Arts and Sciences; 2013.

6. McCulloch, CE, Searle, SR. Generalized, linear, and mixed models. John Wiley & Sons, New York; 2001.

7. Harris, D. Comparison of 1-,2- and 3-parameter IRT models, instructional topics in educational measurement, An NCME Instructional Module on; 1989.

8. Liou, M., Yu, L. 1991. Assesing statistical accuracy in ability estimation: bootstrap approach. *Psychometrika;* 56(1): 55-67.

9. Atanasov, D. 2009. Estimation of IRT parameters over a small sample: Bootstrapping of the item responses. *Pliska Studia Mathematica Bulgaria*; 19: 58-68.

10. Heene, M, Draxler, C, Ziegler, M, Bühner, M. 2011. Performance of the bootstrap Rasch model test under violations of non-intersecting item response functions. *Psychological Test and Assessment Modeling*; 53:283–294.

11. Wolfe, EW, McGill, MT. Comparison of asymptotic and bootstrap item fit indices in identifying misfit to the Rasch model. National Conference on Measurement in Education New Orleans; 2011.

12. Patton, JM, Cheng, Y, Yuan, KH, Diao, Q. 2014. Bootstrap standard errors for maximum likelihood ability estimates when item parameters are unknown. *Educational and Psychological Measurement*; 74(4): 697-712.

13. Olmuş, H., Nazman, E. 2017. An evaluation of the two parameter (2-PL) IRT models through a simulation study. *Gazi University Journal of Science*; 30(1): 235-249.

14. Liu, Y, Yang, JS. 2018. Bootstrap-calibrated interval estimates for latent variable scores in item response theory. *Psychometrika*; 83(2): 333-354.

15. Liu, Y., Hu, G., Cao, L.Wang, X., Chen, M.H. 2019. A comparison of Monte Carlo methods for computing marginal likelihoods of item response theory models. *Journal of the Korean Statistical Society*; 48:503-512.

16. Chen, S., Haziza, D., Leger, C., Mashreghi, Z. 2019. Pseudo-population bootstrap methods for imputed survey data. *Biometrica*; 106(2):369-384.

17. Baker, FB, Kim, SH. Item Response Theory: Parameter Estimation Techniques. Marcel Dekker, Inc; 2004.

18. Partchev, I. 2004. A visual guide to item response theory, Friedrich-Schiller-Universitat Jena. https://www.metheval.uni-jena.de/irt/ VisualIRT.pdf.

19. Hesterberg, T, Monaghan, S, Moore, DS, Clipson, A, Epstein, R. Bootstrap method and permutation tests. W.H. Freeman and Company New York; 2003.

20. Baur, T, Lukes, D. 2009. An Evaluation of the IRT models through monte carlo simulation. *Journal of Undergraduate Research XII:1-7*. Clearinghouse on Assessment and Evaluation.

21. Toribio, SG. Bayesian model checking strategies for dichotomous item response theory models. Graduate College of Bowling Green State University; 2006.