



Comparison of item response theory ability and item parameters according to classical and Bayesian estimation methods

Eray Selçuk ^{1*}, Ergül Demir ²

¹Republic of Türkiye Ministry of National Education, Ankara, Türkiye

²Ankara University, Faculty of Educational Sciences, Department of Educational Measurement and Evaluation, Ankara, Türkiye

ARTICLE HISTORY

Received: May 01, 2023

Accepted: Feb. 13, 2024

Keywords:

IRT parameter estimation,
Maximum likelihood estimation,
Bayesian estimation method,
MCMC,
RMSE.

Abstract: This research aims to compare the ability and item parameter estimations of Item Response Theory according to Maximum likelihood and Bayesian approaches in different Monte Carlo simulation conditions. For this purpose, depending on the changes in the priori distribution type, sample size, test length, and logistics model, the ability and item parameters estimated according to the maximum likelihood and Bayesian method and the differences in the RMSE of these parameters were examined. The priori distribution (normal, left-skewed, right-skewed, leptokurtic, and platykurtic), test length (10, 20, 40), sample size (100, 500, 1000), logistics model (2PL, 3PL). The simulation conditions were performed with 100 replications. Mixed model ANOVA was performed to determine RMSE differentiations. The prior distribution type, test length, and estimation method in the differentiation of ability parameter and RMSE were estimated in 2PL models; the priori distribution type and test length were significant in the differences in the ability parameter and RMSE estimated in the 3PL model. While prior distribution type, sample size, and estimation method created a significant difference in the RMSE of the item discrimination parameter estimated in the 2PL model, none of the conditions created a significant difference in the RMSE of the item difficulty parameter. The priori distribution type, sample size, and estimation method in the item discrimination RMSE were estimated in the 3PL model; the a priori distribution and estimation method created significant differentiation in the RMSE of the lower asymptote parameter. However, none of the conditions significantly changed the RMSE of item difficulty parameters.

1. INTRODUCTION

Test development consists of sequential activities (Thorndike, 1982). Test development processes are carried out within the framework of various theories aimed at minimizing error. In this context, test theories use various methods and models to ensure the reliability and validity of the measurement process. Test theories are an overview that connects observed variables to latent variables. The general purpose of test theories is to estimate the true score. While making this estimation, it is also to determine how much the measurement scores of the defined construct are affected by measurement errors and to find methods to minimize these

*CONTACT: Eray Selçuk ✉ erayselcuk84@gmail.com 📍 Republic of Türkiye Ministry of National Education, Ankara, Türkiye

© The Author(s) 2024. Open Access This article is licensed under a Creative Commons Attribution 4.0 International License. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>

errors. Another purpose of test theory is to help experts become aware of the logical and mathematical models underlying standard practices in test use and construction (Crocker & Algina, 1986).

Two common measurement theories are used in the historical process of the science of psychometrics, which deals with the test development processes and the problems related to their psychometric properties. These are the Classical Test Theory (CTT), which was first developed, and the Item Response Theory (IRT), also called the Latent Trait Theory (LTT), which is increasingly used.

According to IRT, ability or latent trait is performance on test items. IRT is defined as a model that shows the procedure to be followed to establish the consistency between the latent variables and the findings obtained from these variables. IRT should not be seen as a hypothetical theory because this theory does not explain why a person gives an answer to an item or how he/she decides to answer an item. IRT is more of a model based on statistical estimations. IRT uses latent traits of individuals and items to estimate observed responses (Hambleton et al., 1991). In other words, IRT is a statistical theory about how the item under investigation and test performance relates to the abilities measured by the items in the test (Hambleton & Jones, 1993).

The advantages of IRT models can be achieved only when the fit between the model and test data is satisfactory (De Mars, 2010). The most important conditions for ensuring this harmony are appropriate sample size, adequate test length, and a normal priori distribution type. These conditions significantly affect the amount of error, especially in parameter estimation. In addition, although the number of standard error rates of parameter estimations depends on sample size and test length, estimation methods also affect this amount of standard error. In addition, there are some assumptions that estimation methods can work effectively. In terms of data, if these assumptions are ignored and neglected, the error rates in the estimations increase (Hambleton & Swaminathan, 1985).

There are different methods for estimating item and ability (person) parameters within the framework of IRT. Most of these methods are based on calculating the maximum likelihood (ML) function. The ML function is calculated by estimating the probabilities of the values, maximizing the item and ability parameters over the observed data. These estimation methods perform a solution with an iterative process. The most critical limitation of ML functions, in general, is that it is not possible to estimate the ability parameters of individuals with a full or zero score on a test or to estimate the parameters of the items that are correctly or incorrectly made by everyone (Lord, 1983; Samejima, 1993a, 1993b). In addition, the priori distribution type (normal, skewed left, skewed right, leptokurtic, and platykurtic) effectively estimates item and ability parameters and determines the standard errors of these estimations. In cases where the distribution becomes skewed or when the aforementioned general problems of the ML methods are encountered, "Bayesian Estimation Methods" are recommended to make estimations meticulously (with a lower standard error rate) (Bock & Mislevy, 1982; De Ayala, 2009; Hambleton & Swaminathan, 1985; Hambleton et al., 1991).

ML methods cannot accurately estimate item and ability parameters in generally small samples, short tests, and especially in skewed data. Likewise, the increase in the number of parameters in the IRT model (as in the 2 PL and 3 PL models) also increases the error in these estimations. The literature recommends parameter estimation with the Bayesian approach for such problems.

Most likelihood methods used in IRT are based on the frequency approach. However, the frequency approach has shortcomings because it depends on a fixed value and does not provide distribution information. The Bayesian approach allows estimations by including a priori distribution information. In the Bayesian approach, the variance of the prior distribution

represents the uncertainties of the parameter estimates. If the variance of the prior distribution is low, the error rates of the parameter estimates will be lower (van de Schoot & Depaoli, 2014). Using a Bayesian approach will solve some of the difficulties encountered with the ML approach. Bayesian estimates for the level of ability (Θ) can be obtained for zero correct response item patterns, fully correct response item patterns, and anomaly response patterns (Hambleton et al., 1991).

Bayesian IRT estimation methods can provide advantages over ML IRT estimation methods (Bock & Mislevy, 1982; De Ayala, 2009; Hambleton & Swaminathan, 1985; Hambleton et al., 1991). The essence of the Bayesian approach is to know the individual's point in the distribution in terms of a trait before obtaining any data. This distribution is called a priori distribution. Therefore, restricting parameter estimations to specific ranges using a priori distribution is essential for Bayesian estimations of IRT (Gao & Chen, 2005).

Gao and Chen (2005) conducted a large-scale simulation study on 3 PL models. In their study, authors used uniform distribution data sets with test lengths of 10, 20, and 60 items and sample sizes of 100, 200, and 500. The authors compared the marginal maximum likelihood (MML) estimation method and Bayesian estimation methods on these data. As a result of the research, the authors concluded that the marginal maximum likelihood method tends to estimate out of the true item parameter values in small samples. Moreover, the authors stated that Bayesian estimation yielded more accurate estimates than marginal maximum likelihood estimation when the sample size was as low as 100. The authors emphasized that the results of Bayesian estimation are more satisfactory regarding the root mean standard error of the estimates (RMSE). However, the error amounts of the marginal maximum likelihood estimation methods also tend to decrease when the test length and sample size increase.

Sass et al. (2005) compared the estimation errors of the latent trait distribution under normal and non-normal distributions. The authors simulatively generated data for 1000 samples, 30 items, and 2 PL models. They used maximum likelihood (ML), Bayesian MAP, and EAP as parameter estimation methods. They also examined true and estimated item parameters to distinguish item parameter estimation from latent trait estimation errors. They stated that non-normal latent trait distributions produce higher estimation errors than normal distributions.

Accordingly, while estimating the parameters based on IRT, the data are the problem of this research is whether there will be a difference between the RMSE of the estimations when the priori distribution type is manipulated in terms of sample size, test length, and logistics model compared to ML and Bayesian IRT. For this purpose, answers to the following research problems were sought through the data generating according to simulation conditions:

1. Is there a significant difference between the RMSE of the ability parameters (Θ_{RMSE}) estimated by ML and Bayesian methods in the generated datasets in 2 PL models according to simulation conditions?
2. Is there a significant difference between the RMSE of the ability parameters (Θ_{RMSE}) estimated by ML and Bayesian methods in the generated datasets in 3 PL models according to simulation conditions?
3. Is there a significant difference between the RMSE of item discrimination (a_{RMSE}), RMSE of item difficulty (b_{RMSE}) and RMSE of lower asymptote (c_{RMSE}) estimated by ML and Bayesian methods in the generated datasets in 2 PL models according to simulation conditions?
4. Is there a significant difference between the RMSE of item discrimination (a_{RMSE}), RMSE of item difficulty (b_{RMSE}) and RMSE of lower asymptote (c_{RMSE}) estimated by ML and Bayesian methods in the generated datasets in 3 PL models according to simulation conditions?

Estimation methods are affected by the distributional types of persons' abilities and item parameters. It is also assumed that most traits (Θ) are normally distributed in the universe. This assumption reveals the strengths of IRT and affects the estimation of parameters. Therefore, skewed distributions cause some issues in parameter estimation. This is because the accurate

estimation of parameters depends on the variance not being sufficiently large at some levels of Θ . If such distributional assumptions are not satisfied, the accuracy of parameter estimation based on maximum likelihood methods of IRT is questionable. In conclusion, this research is essential in the sense that it acknowledges that parameters estimated with different a priori ability distributions other than the normal distribution (left and right skewed, leptokurtic and platykurtic) have high RMSE and proposes an alternative estimation method to reduce this error and Bayesian approach provides advantages in parameter estimation compared to the ML approach.

1.1. Significance of the Research

The studies by Swaminathan and Gifford (1986), Harwell and Janosky (1991), Gao and Chen (2005), Sass et al. (2005), Finch and Edwards (2015), Çelikten and Çakan (2019) and Kıbrıslıoğlu Uysal (2020) compared different estimation methods on IRT parameter estimation under different conditions. It is seen that most comparisons were made under the conditions of sample size and test length, and the most used estimation methods were likelihood (ML), MAP, and EAP. Studies also investigate the effect of latent trait or item parameter distributions. These studies were generally conducted on simulative data.

This research aims to compare different sample sizes, test lengths, latent trait distributions, and parameter estimation methods with the effect of manipulating conditions as in the previous studies. The research is similar to other studies in this respect. However, the distinguishing feature of this research is that five different types of a priori ability distributions were generated; accordingly, the total test scores also had this distribution type. However, there are some studies in which the latent distribution is skewed. This study analyzed skewness as bidirectional (left-skewed and right-skewed), and leptokurtic and platykurtic distributions were also examined. In addition, in some previous studies, Bayesian estimation has usually been analyzed in Rasch or 2 PL models. This research also examined the results of Bayesian MCMC parameter estimation in the 3 PL model.

As a result of the research, it is foreseen that using Bayesian estimation methods in situations where sample size and test length are not enough for a priori distributions of ability in different patterns will lead to low RMSE in parameter (ability and item) estimations. From this point of view, this research is thought to provide a different viewpoint on the parameter estimation methods used in IRT and contribute to the literature.

2. METHOD

2.1. Research Design

This research created data sets with different the priori distribution types following the simulation conditions. Estimations of ability and item parameters were made using ML and Bayesian (MCMC) methods on these data sets. Simulation studies can use data generated in simulative conditions to investigate certain variables. The simulation approach creates an artificial condition where relevant information and data can be generated. This enables us to observe the dynamic behavior of a system (or sub-system) under controlled conditions (Fraenkel & Wallen, 2009; Kothari, 2004). The literature argues that simulation studies are empirical experiments (Morris et al., 2017) and should be considered statistical sampling, depending on the research design and data analysis principles determined (Hoaglin & Andrews, 1975). Accordingly, this research uses a statistically experimental method to compare estimation methods by manipulating various conditions through simulatively generated data. In this respect, this research is a simulation-based experimental study.

2.2. Generating Data

Monte Carlo (MC) simulation generates the data within the scope of this study following the conditions manipulated in different ways according to the prior distribution types, sample size, test length, logistics model and parameter estimation method specified in the research problem.

Monte Carlo (MC) simulation is used in many applications, such as evaluating new methods in IRT parameter estimation, performance comparison of different item analysis programs, and parameter estimation in multidimensional data. Accordingly, IRT applications using the MC simulation technique should include at least one of the following (Harwell et al., 1996):

1. Evaluation of parameter recovery or parameter estimation methods,
2. Evaluation of the properties of IRT-based statistics,
3. Methodological comparison by combining different IRT applications.

The R programming language generated the data depending on the simulation conditions. In R, *mirt* (Chalmers, 2012), *e1071* (Meyer, 2022), *psych* (Revelle, 2022) and *lattice* (Sarkar, 2022) packages were run. The *simdata* function in the *mirt* package generated binary (1-0) score matrices with the "Önsel (Prior)" script block written by the researchers, according to the simulation conditions. The "Önsel (Prior)" script block is given in [Appendix](#). While generating the binary score matrices, the priori ability scores produced by the distribution types were placed in the latent distribution argument within the *simdata* function.

In generating the data in the "Önsel (Prior)" script block, previous research in the literature was referred to for the initial item parameters. Accordingly, log-normal distribution [$a \sim \ln N(0.3, 0.2)$] was used to generate the item discrimination parameter, standard normal distribution [$b \sim N(0, 1)$] was used to generate the item difficulty parameter, and uniform distribution [$c \sim U(0.01, 0.25)$] was used to generate the item chance parameter (lower asymptote) (Baker, 2001; Feinberg & Rubright, 2016; Bulut & Sünbül, 2017; Soysal, 2017; Pekmezci, 2018). In generating the a priori ability parameter, more than one and different (normal and uniform) distribution types were combined. In the generation of skewed, leptokurtic, and platykurtic distributions other than the normal distribution, outliers were generated at Z scores above ± 4 . Accordingly, $\Theta \sim N(0, 1)$ if the distribution is normal; $\Theta \sim N(2, 1)$, $\Theta \sim U(-5.0, -4.0)$ and $\Theta \sim U(-4.0, -3.0)$ if the distribution is left skewed; $\Theta \sim N(-2, 1)$, $\Theta \sim U(3.0, 4.0)$ and $\Theta \sim U(4.0, 5.0)$; $\Theta \sim N(-1, 100)$, $\Theta \sim N(1, 100)$ and $\Theta \sim N(0, 0.00001)$ if leptokurtic; $\Theta \sim N(0, 1)$, $\Theta \sim U(-3.0, -1.0)$ and $\Theta \sim U(1.0, 3.0)$ if platykurtic.

Considering skewed distributions with normal distribution assumptions leads to incorrect results (Kolen, 1985). Deviations from the normal distribution cause various problems when estimating parameters with ML estimation methods (Hambleton & Swaminathan, 1985). For this reason, the problem of this research is how different a priori ability distribution types will affect parameter estimation methods.

2.3. Simulation Conditions

In the simulation model created to solve the problems in this research, some conditions were fixed while others were manipulated. According to the literature, the selection of each condition in the research was determined by examining similar previous studies. The conditions that were fixed and manipulated are given in [Table 1](#).

Table 1. Conditions of simulation.

Conditions of Simulation							
Fixed conditions				Manipulated conditions			
Model Parameters		Parameter estimation methods (x2)		Sample size (x3)	Test length (x3)	Logistics model (x2)	Prior distribution type (x5)
Initial of ability parameters (Θ_i)	Initial of item parameters (a_i, b_i, c_i)	Maximum likelihood (ML)	Bayesian (MCMC)	100	10	2 PL 3 PL	Normal
				500	20		Left-skewed
				1000	40		Right-skewed
							Leptokurtic
							Platykurtic

Table 1 shows that the research conditions consist of fixed and manipulated conditions. Fixed conditions, initial of model parameters, and manipulated conditions were determined as estimation method, sample size, test length, logistics model, and priori distribution type. Accordingly, parameter estimation methods (ML x Bayesian), sample size (100 x 500 x 1000), test length (10 x 20 x 40), logistics model (2 PL x 3 PL), and priori distribution type (normal x left-skewed x right-skewed x leptokurtic x platykurtic) 180 simulation conditions were carried out with 100 replications. Accordingly, 18000 data sets were used in the research process.

Determining the simulation conditions is essential in reviewing previous research in the literature and determining which factors should be selected to contribute to the literature. In the simulation model developed to solve the research problems in this study, some conditions were kept fixed while others were manipulated.

2.3.1. Fixed conditions

Model parameters (ability and item parameters): The initial parameters used to generate ability and item parameters are given in the data generation section.

2.3.2. Manipulated conditions

Parameter estimation method: Maximum likelihood (ML) and Bayesian MCMC methods were used to estimate the ability and item parameters. These estimation methods were used for each simulation condition and replications separately. Moreover, this condition is one of the most critical problems the research aims to address.

Sample size: For each simulation condition, three different sample sizes of 100, 500, and 1000 participants were selected. Sample size is considered an essential variable for IRT estimation (Hambleton, 1989; Orlando, 2004). The strengths of IRT depend on the sample size, and it is suggested that it should be applied in large samples (DeMars, 2010). Linacre (1994) stated that small samples are needed when the number of parameters in the model is less, while more complicated models need larger samples. In the literature, there are some studies indicating that sample sizes of 200 (Wright & Stone, 1979) or 500 (Hulin et al., 1982) for 1 PL model, 1000 (Ree & Jensen, 1980) for 2 PL model, and 1000 (Lord, 1968) or 10000 or more (Thissen & Wainer, 1983) for 3 PL model are adequate. In addition, De Ayala (2009) stated that sample sizes of 250 or 500 are adequate for parameter estimation, whereas Hulin et al. (1982) concluded that a sample size of more than 2000 is unnecessary for parameter estimation using ML methods in general. Mislevy (1986) used a sample of 1000 in his study on parameter estimation using Bayesian approach. In this study, we want to utilize the advantages of Bayesian approach by using different sample sizes. Therefore, data sets of 100 for a small sample size, 500 for a medium sample size, and 1000 for a large sample size were used.

Test Length: Three different test lengths were selected for each simulation condition: 10, 20, and 40 items. Using different test lengths leads to a variation in the item response patterns. This variation is especially crucial for the accuracy of item parameter estimates (Hulin et al., 1982). As the test length increases, the accuracy of Θ estimations increases. Accordingly, increasing the sample size and test length will increase the accuracy of the estimation item parameters (a_i , b_i , and c_i) and thus increase the accuracy of the ability parameter (Θ) estimates (Reise & Yu, 1990). DeMars (2010) stated that for 2 PL and 3 PL models, the test length should be 20 when using a sample of 500, 40 items when using a sample of 1000, and 50 to 80 items when using a sample of 2000-3000. Hulin et al. (1982) suggest that using a 30-item test in a sample of 500 in 2 PL models and a 60-item test in a sample of 1000 in 3 PL models would be adequate in terms of the accuracy of parameter estimations. Hambleton and Cook (1983) stated that a 20-item test in a sample of 500 in the 3 PL model is adequate for parameter estimation. However, Hambleton and Cook (1983) stated that the estimation error was negatively affected when the test length increased to 40. Akour and Al-Omari (2013) stated that a test length of 15 items in a sample of 200 is sufficient for parameter estimation in the 3 PL model. Mislevy (1986) used 20 and 40 items as test lengths in his study on parameter estimation with the Bayesian approach.

This study generated data sets of 10 items for short tests, 20 for medium length tests, and 40 for longer tests. Although short tests are mostly teacher-made tests in classroom assessments, these tests are now also used in secondary education entrance examinations in Turkey. In these examinations, the number of items in the Turkish History of Turkish Revolution and Kemalism subtests, Religious Culture and Ethics, and Foreign Language, is 10 (MoNE LGS Guide, 2022). For this purpose, 10 items were selected as test length, one of the simulation conditions.

Priori Distribution of Ability (Theta): Each simulation condition used five different types of distributions, keeping the standard deviation values fixed. The simulation conditions were selected as normal and non-normal (left-skewed, right-skewed, leptokurtic, and platykurtic) distribution types. The skewness coefficient's absolute value means that the samples' distribution types are highly skewed when greater than 1.00, moderately skewed between 0.50 and 1.00, and approximately symmetric when less than 0.50. For kurtosis, it is stated that the distribution is normal if the coefficient is 3, leptokurtic if it is greater than 3, and platykurtic if it is less than 3 (Bulmer, 1979). However, with the addition of -3 to the formula, this value becomes 0. This means that a kurtosis coefficient of 0 indicates that the distribution is normal, a coefficient greater than 0 indicates that the distribution is leptokurtic, and a coefficient less than 0 indicates that the distribution is platykurtic. Tabachnick and Fidell (2014) stated that when the skewness and kurtosis values are between -1.50 and +1.50, the distribution is assumed to be normal. Evaluating skewed distributions with normal distribution assumptions causes incorrect conclusions (Kolen, 1985). It is known that deviations from the normal distribution cause various problems when estimating parameters with maximum likelihood estimation methods (Hambleton & Swaminathan, 1985). For this reason, the issue of this study is how different a priori ability distribution types will affect parameter estimation methods.

IRT Model: This research selected 2 PL and 3 PL models for parameter estimations. According to Hulin et al. (1982), these logistic models are robust and the most widely used models.

Accordingly, two different references were considered when setting the simulation conditions. The first one is to benefit from similar studies in the literature while setting each condition, and the second one is to consider the advantages of the Bayesian estimation method depending on the purpose of the research. In the first reference, the previous research related to the literature is discussed in detail under the topic of each condition. In the second reference, these conditions were selected by considering the problems of ML estimation and the advantages of Bayesian estimation. Since this study aims to determine how the ML and Bayesian estimation results will change, especially in cases where the sample becomes smaller, the number of items decreases. The prior ability distribution becomes skewed; this is another significant reason for choosing the simulation conditions in this way.

Harwell et al. (1996) suggested that at least 25 replications should be used in studies where the IRT parameters are manipulated. However, Seong (1990) used 5 replications, Stone (1992) used 100 replications, Kirisci et al. (2001) used 10 replications, Sass et al. (2008) used 100 replications, Finch and Edwards (2015) used 1000 replications, Bulut and Sünbül (2017) used 100 replications, Karadavut (2019) used 25 replications, and Kıbrıslıoğlu Uysal (2020) used 100 replications in various simulation studies given in related studies.

A literature review shows that similar simulation studies use different numbers of replications when generating data. There are two factors affecting this issue. The first is that the degree of accuracy of the data generated because of a low number of replications is insufficient, and the second is that the simulation program is inadequate and time costly because of many replications (Bulut & Sünbül, 2017). Moreover, Feinberg and Rubright (2016) proposed a formulation for the number of replications in IRT simulations based on the standard deviation of the estimated parameters. This equation is given below:

$$\sigma_M = \frac{\hat{\sigma}}{\sqrt{R-1}} \quad (\text{Equation 1})$$

where σ is the standard deviation of the estimated parameter across replications, R is the number of replications, and σM is the standard error of the mean. Accordingly, researchers determine an initial number of replications, and after computing the standard deviation of the data, they set a new number of replications. If the standard deviation is larger than expected Feinberg and Rubright (2016) recommend increasing the number of replications. However, there is no acceptable value for the estimated standard deviation value. Therefore, Barış-Pekmezci and Şengül-Avşar (2021) state that it is not practical to use this equation. Therefore, considering the research previously cited in the literature, it was decided to use 100 replications in this study to produce accurate results and not to increase the simulation time.

2.4. Analysis of Data

First, basic assumptions were checked to determine the fit of the generated datasets for IRT parameter estimation. These assumptions are unidimensionality, local independence, and model-data fit (Baker, 2001; Baker & Kim, 2004; De Ayala, 2009; Hambleton & Swaminathan, 1985; Hambleton et al., 1991; Hambleton & Jones, 1993). The psych (Revelle, 2022), sirt (Robitzsch, 2022), and mirt (Chalmers, 2012) packages were used to test the basic assumptions.

Second, the R programming language was used in the analysis of the data as well as in the generating of the data. The R software version used is R Studio, Version: 2022.12.0+353. Researchers generally use statistics such as correlation, covariance, bias, absolute bias, standard error of estimate (SE), mean square error (MSE) and root mean square error (RMSE). The statistics to be used and how to interpret them depend on the problems of the research. A review of the literature shows that bias, standard error (SE) and root mean square error of the mean square error (RMSE) are the most used ones (Feinberg & Rubright, 2016). RMSE was used in this research.

Root means square error (RMSE) between the ability and item parameters and the initial parameters estimated on the data generated according to the simulation conditions were calculated. This is because biased values can take both positive and negative values. This situation affects the mean of bias. In addition, there is a relationship between RMSE and bias. This relationship is given in the equation below (Atar, 2007; Bilir, 2009; Feinberg & Rubright, 2016):

$$RMSE^2 = Bias^2 + SE^2 \quad (\text{Equation 2})$$

In this equation, the sum of the bias's square and the standard error's square equals the square of the RMSE. Accordingly, the negative and positive biases created by the bias have disappeared. While analyzing the data, the ML estimation was first performed using the irtplay package (Lim & Wells, 2020) compared to ML approaches, followed by standard Bayesian estimations using Monte Carlo Markov Chain (MCMC) methods using the bairt (Martinez, 2017) and sirt (Robitzsch, 2022) packages for Bayesian approaches. In Bayesian estimations, the burning was defined as 1000, and the iteration was defined as 3000. The number of burn-in and iterations are set at these values due to the procedures performed in the algorithm of the method. Because in the MCMC method, the first chain generated up to the burn-in value is subtracted from the whole chain generated later. Thus, parameter estimation is performed from the sample generated by the number of iterations (Martin & Quinn, 2006; SAS Institute, 2020). These values are determined according to the conditions of the simulation to provide unbiased results at the expected level.

Third, the significance of the differences between the RMSE values of the parameters was tested by mixed model ANOVA according to sample size, test length, logistic model, a priori distribution type, and estimation method. Assumptions were checked before analyzing the mixed model ANOVA. Afex (Singmann, 2022) and emmeans (Lenth, 2022) packages were used for this analysis. For the mixed model ANOVA, the main effects (between) variables were the simulation conditions that were manipulated (sample size, test length, a priori ability distribution types, parameter estimation methods) and fixed (initial values of ability and item

parameters), and the number of simulation replications was assigned as the interaction (within) variable. According to the analysis results, the significant conditions' effect sizes (generalized eta-square coefficient) were computed and assessed according to Cohen's (1988) proposal. Accordingly, the size of the effect size was interpreted as weak if it was less than 0.0099, moderate if it was 0.0588, and strong if it was greater than 0.1379. At the same time, since the generalized eta-square coefficient takes a value between 0 and 1 when this value is multiplied by 100, it shows how much of the variance of the dependent variable is explained by the independent variables (Lakens, 2013). Statistically significant conditions were compared using the Bonferroni post hoc comparison method, included by default in the emmans package. According to the analysis results, ggplot2 (Wickham, 2016) and ggbeeswarm (Clarke, 2022) packages were used to visualize significant conditions.

3. RESULTS

Analysis was conducted to determine whether the datasets meet the assumptions of the IRT. Accordingly, for the unidimensionality assumption, the ratio of the explained variance, the averages of the first eigenvalues and the ratio of the first eigenvalue to the second eigenvalue were calculated according to the explanatory factor analysis results. It was accepted that this assumption was fulfilled if a dominant factor was found (Lord, 1980). Accordingly, it is seen that the data fulfills the unidimensionality assumption in all conditions.

The Q3 statistic of Yen (1984) is used to test the local independence assumption. Accordingly, it is determined that the local independence assumption is mostly fulfilled for the data in all conditions.

M2 values were examined to test the assumption of model-data fit. As a fit criterion, the M2 statistic is expected to be non-significant (Maydeu-Olivares & Joe, 2006). Accordingly, it is seen that model-data fit is fulfilled in all the data.

Normality and homogeneity of variances test results of the data were analyzed. In big samples, it is more practical to use descriptive statistics and graphical analysis to check the normality assumption. In big samples, normality tests with hypothesis tests risk increasing the probability of Type I error (Demir, 2019). Accordingly, it is seen that the skewness and kurtosis coefficients and histogram graphs of the data fulfill the normality assumption. Examining the hypothesis of homogeneity of variances test results shows that this assumption is fulfilled ($F_{(2PL.0.RMSE)} = 0.13$; $p > .05$, $F_{(2PL.a.RMSE)} = 0.51$; $p > .05$, $F_{(2PL.b.RMSE)} = 0.78$; $p > .05$, $F_{(3PL.0.RMSE)} = 0.06$; $p > .05$, $F_{(3PL.a.RMSE)} = 0.21$; $p > .05$, $F_{(3PL.b.RMSE)} = 0.99$; $p > .05$, $F_{(3PL.c.RMSE)} = 0.59$; $p > .05$). Then, the findings related to the research problems are presented under headings.

3.1. Investigation of Θ_{RMSE} Estimated by ML and Bayesian Methods in 2 PL Model

In the first problem of the study, the RMSE changes of ability parameters according to sample size, test length, and estimation method were analyzed with mixed model ANOVA in the data in the 2 PL model with normal and non-normal priori distribution (left-skewed, right-skewed, leptokurtic and platykurtic). Accordingly, the results of the mixed model ANOVA performed for the ability parameters according to the sample size, test length, and estimation method in the data in the 2 PL model with normal and non-normal priori distribution (left-skewed, right-skewed, leptokurtic, and platykurtic) are given in [Table 2](#).

Table 2. Mixed model ANOVA results for ability parameters RMSE in data in 2 PL models with normal and non-normal priori distribution.

Independent variables	Mean squares of error	Degrees of freedom	F	p	Generalized η^2
Estimation method (K)	72.05	1	7.83	0.006**	0.078
Sample size (S)	79.36	2	0.00	0.997	0.001
Test length (M)	65.24	2	9.42	0.001**	0.171
Prior distribution type (D)	42.74	4	1.88	0.001**	0.456
K*S	75.46	2	0.01	0.994	0.001
K*M	58.50	2	1.69	0.191	0.037
K*D	31.90	4	4.05	0.005**	0.155
Error	0.30	198			
Total	425.55				

* $p < .05$, ** $p < .01$

Table 2 shows that the main effects of the estimation method ($F_{(1, 88)} = 7.83$; $p < .01$, $\eta^2 = .078$), test length ($F_{(2, 84)} = 9.42$; $p < .01$, $\eta^2 = .171$) and priori distribution type ($F_{(4, 80)} = 1.88$; $p < .01$, $\eta^2 = .456$) seem to have a significant effect. However, the sample size ($F_{(2, 87)} = 0.00$; $p > .05$, $\eta^2 = .001$) did not have a significant effect. Significantly, the estimation method has a medium effect size, the test length is high, and the priori distribution type has a high effect size. When the interactions were examined, the interaction between the estimation method and the priori distribution type was significant ($F_{(4, 80)} = 4.05$; $p < .01$, $\eta^2 = .155$). The effect size of the interaction is high. Pairwise comparisons of the ability parameter estimation method in 2 PL models are given in Table 3.

Table 3. Ability parameter estimation method pair comparisons in 2 PL models.

Estimation method	Difference	Standard error	t	p
Bayes-ML	-0.501	0.179	-2.799	0.001**

* $p < .05$, ** $p < .01$

Table 3 shows that Bayesian estimation, the ability parameter estimation method in the 2 PL model, produced lower and more significant RMSE than the ML ($t = -2.799$; $p < .01$). The RMSE changes of the ability parameter estimation methods in the 2 PL model are given in Figure 1.

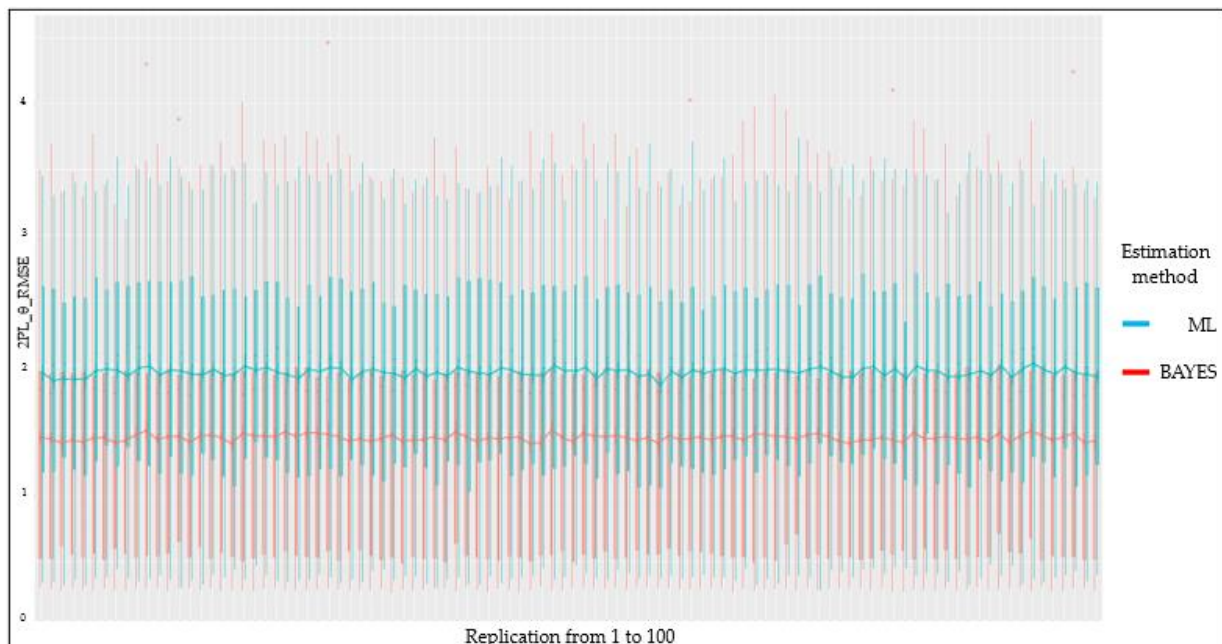
Figure 1. The change of ability parameter RMSE in 2 PL models by estimation methods.

Figure 1 shows that the RMSE of the ability parameters obtained from all data sets in the 2 PL model, regardless of the research conditions, change. Accordingly, while the ability parameter was estimated in the 2 PL model, the Bayesian method produced lower and more significant RMSE than the ML method. Pairwise comparisons according to the number of items on the ability parameter in the 2 PL model are given in Table 4.

Table 4. Pairwise comparisons of ability parameter RMSE in 2 PL models by test length.

Test Length	Difference	Standard error	<i>t</i>	<i>p</i>
10 – 20	0.272		1.304	0.397
10 – 40	0.884	0.209	4.236	0.001**
20 – 40	0.612		2.933	0.012*

* $p < .05$, ** $p < .01$

Table 4 shows that there are significant differences between test lengths 10 and 40 ($t=4.236$; $p<.01$) and 20 and 40 ($t=2.933$; $p<.05$) on ability parameter RMSE in the 2 PL model. The RMSE change according to test length on the ability parameter in the 2 PL model is given in Figure 2.

Figure 2. The change of ability parameter RMSE in 2 PL models by the test length.

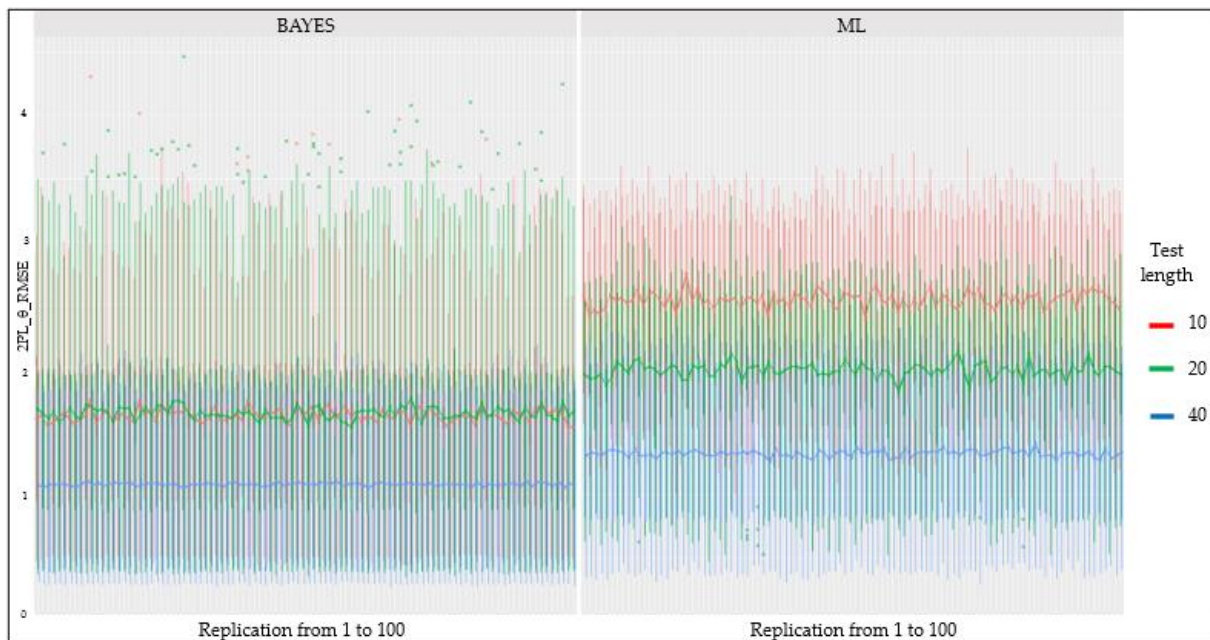


Figure 2 shows that RMSE decreases as the test length increases on the ability parameter estimations in the 2 PL model. As a result of the estimation made with the ML, the RMSE of the ability parameters decreases as the test length increases. The same situation is seen in the Bayesian estimation method. In the Bayesian estimation method, there is no difference in the test length between 10 and 20, but a lower RMSE is obtained in case the test length is 40. However, the RMSE of ability parameters obtained according to test length in Bayesian estimation was lower than in ML estimation. Pairwise comparisons according to priori distribution on the ability parameter in the 2 PL model are given in Table 5.

Table 5 shows that the priori distribution type on the ability parameter RMSE in the 2 PL model is normal to left skewed ($t=-7.292$; $p<.01$), normal to right skewed ($t=-7.321$; $p<.01$), normal to leptokurtic ($t=-5.434$; $p<.01$), normal to platykurtic ($t=-3.267$; $p<.05$), left skewed to platykurtic ($t=4.026$; $p<.01$), right skewed to platykurtic ($t=4.054$; $p<.01$) significant differences were found. These differences are in favor of the Bayesian estimation method. In the 2 PL model, Bayesian estimation produces lower RMSE as the priori distribution type differs from the

normal. The RMSE change according to the priori distribution type on the ability parameter in the 2 PL model is given in Figure 3.

Table 5. Pairwise comparisons of ability parameter RMSE in 2 PL models by prior distribution.

Prior talent distribution type	Difference	Standard error	<i>t</i>	<i>p</i>
Normal – Left skewed	-1.589		-7.292	0.000**
Normal – Right skewed	-1.595		-7.321	0.000**
Normal – Leptokurtic	-1.184		-5.434	0.000**
Normal – Platykurtic	-0.712		-3.267	0.013*
Left skewed – Right skewed	-0.006	0.218	-0.029	0.999
Left skewed – Leptokurtic	0.405		1.858	0.347
Left skewed – Platykurtic	0.877		4.026	0.001**
Right skewed – Leptokurtic	0.411		1.887	0.332
Right skewed – Platykurtic	0.884		4.054	0.001**
Leptokurtic – Platykurtic	0.472		2.267	0.202

* $p < .05$, ** $p < .01$

Figure 3. The change of ability parameter RMSE in 2 PL models by prior distribution type.

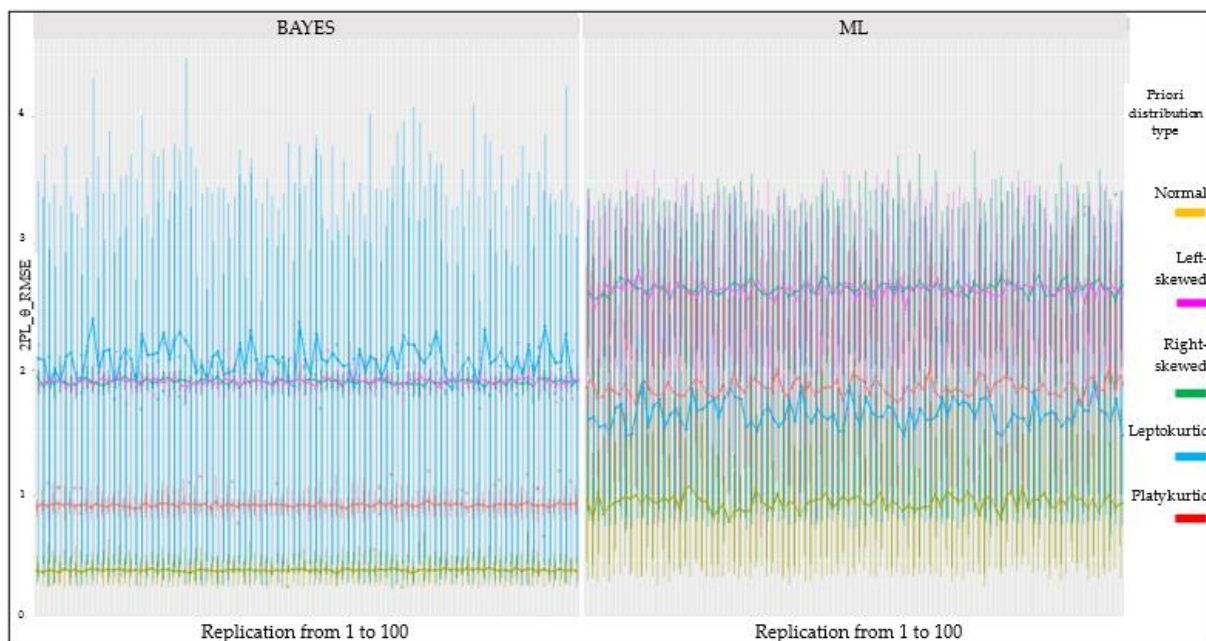


Figure 3 shows that the priori distribution in the 2 PL model becomes skewed from normal (left skewed, right skewed, leptokurtic, and platykurtic), and the RMSE of the ability parameters increases in the ML estimation. The lowest RMSE on the ability parameters was obtained in ML estimation when the prior distribution was normal. As the distribution becomes skewed, the error values increase. The RMSE is highest when the distribution is left skewed and right skewed and lower when it is leptokurtic and platykurtic. As the distribution normalizes, these values show a further decrease. In the 2 PL model, when the Bayesian method performs the ability parameters estimation, RMSE is lower than the ML estimation.

Similarly, the lowest RMSE is in the normal, platykurtic, left and right skewed distribution and the leptokurtic distribution, respectively. In all the priori distribution types, except for the leptokurtic distribution, the RMSE decreases in Bayesian estimation. In contrast, in the leptokurtic distribution, they have higher values than the ML estimation. When the prior distribution is produced, since the leptokurtic distribution has a lower standard deviation than the normal distribution and remains relatively between -1 and +1 as a distribution range, it takes shape in a broader range as a posterior distribution compared to the prior distribution. Therefore, the RMSE differences between the initial and estimated ability parameters increase.

Accordingly, while estimating the ability parameters in the 2 PL model, using the Bayesian estimation method in other distribution types provides lower RMSE, except when the priori distribution is leptokurtic.

3.2. Investigation of Θ_{RMSE} Estimated by ML and Bayesian Methods in 3 PL Model

In the second problem of the study, the RMSE changes of ability parameters according to sample size, test length, and estimation method were analyzed with mixed model ANOVA in the data in the 3 PL model with normal and non-normal priori distribution (left-skewed, right-skewed, leptokurtic and platykurtic). Accordingly, the mixed model ANOVA results were performed for the ability parameters according to the sample size, test length, and estimation method in the data in the 3 PL model with normal and non-normal priori distribution (left-skewed, right-skewed, leptokurtic, and platykurtic) are given in Table 6.

Table 6. Mixed model ANOVA results for ability parameters RMSE in the data in the 3 PL model with normal and non-normal priori distribution.

Independent variables	Mean squares of error	Degrees of freedom	F	<i>p</i>	Generalized η^2
Estimation method (K)	2769.62	1	0.27	0.607	0.003
Sample size (S)	2747.44	2	0.99	0.376	0.022
Test length (M)	2488.40	2	5.62	0.005**	0.111
Priori distribution type (D)	2315.05	4	5.15	0.001**	0.189
K*S	2836.73	2	0.00	0.999	0.001
K*M	2568.23	2	0.00	0.996	0.001
K*D	2441.93	4	0.07	0.991	0.003
Error	0.96	198			
Total	18168.36				

p* < .05, *p* < .01

Table 6 shows that the test length is the main effect of the independent variables ($F_{(2, 84)} = 5.62$; $p < .01$, $\eta^2 = .111$) according to the mixed model ANOVA results for the ability parameters RMSE in the data in the 3 PL model with normal and non-normal priori distribution and priori distribution type ($F_{(4, 80)} = 5.15$; $p < .01$, $\eta^2 = .189$) were found to be significant. The estimation method ($F_{(1, 88)} = 0.27$; $p > .05$, $\eta^2 = .003$) and sample size ($F_{(2, 87)} = 0.99$; $p > .05$, $\eta^2 = .022$) do not have a significant difference. Significantly, the test length is medium, and the priori distribution type has a high effect size. When the interactions were examined, no condition was found to be significant. Pairwise comparisons according to the test length on the ability parameter in the 3 PL model are given in Table 7.

Table 7. Pairwise comparisons of ability parameter RMSE in 3 PL models by test length.

Test length	Difference	Standard error	<i>t</i>	<i>p</i>
10 – 20	3.429		2.663	0.025*
10 – 40	3.988	1.288	3.096	0.007**
20 – 40	0.558		0.434	0.902

p* < .05, *p* < .01

Table 7 shows that there are significant differences between test lengths 10 and 20 ($t=2.663$; $p < .05$) and 10 and 40 ($t=3.096$; $p < .01$) on ability parameter RMSE in the 3 PL model. The RMSE change according to test length on the ability parameter in the 3 PL model is given in Figure 4.

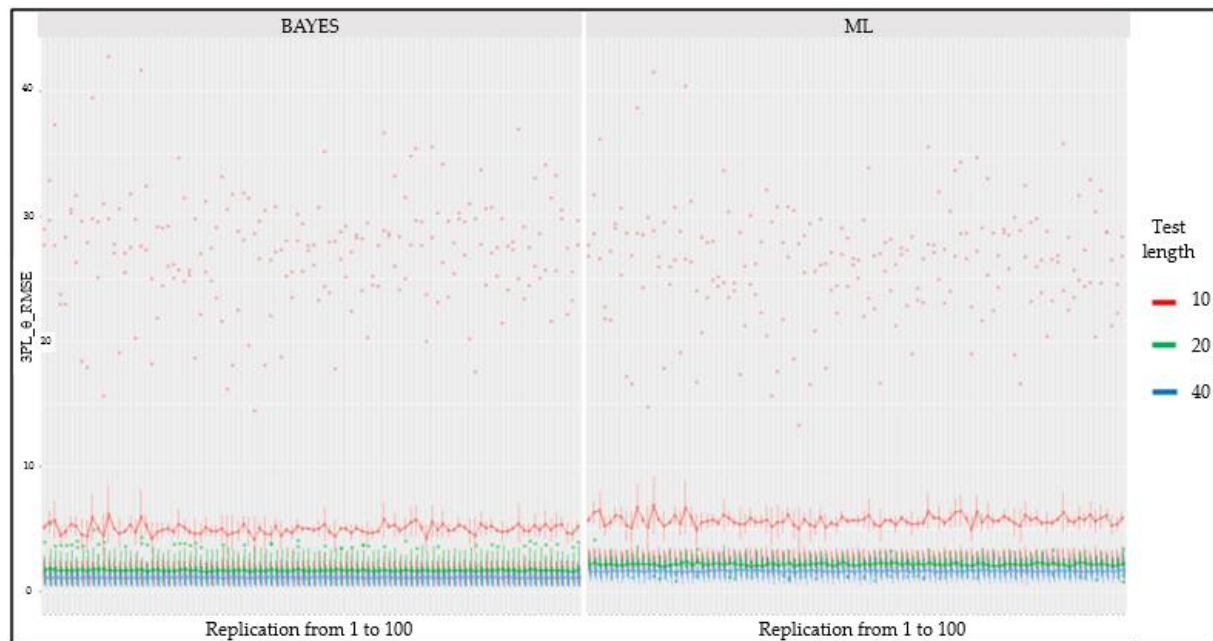
Figure 4. The change of ability parameter RMSE in 3 PL model by estimation methods.

Figure 4 shows that the RMSE decreases as the test length increases on the ability parameter estimations in the 3 PL model. At the same time, the Bayesian estimation method took lower values than ML estimation in cases where test length decreased. However, this situation was not found to be significant. Therefore, using ML or Bayesian methods does not make a difference when estimating ability parameters in the 3 PL model. However, regardless of the estimation method used, the increase in test length causes a decrease in the RMSE of the ability parameters. For example, when the test length decreased to 10, RMSE in the ability parameters increased significantly. Accordingly, lower RMSE for ability parameters in the 3 PL model was observed when the test length was 20 and 40. Pairwise comparisons according to prior distribution type on the ability parameter in the 3 PL model are given in Table 8.

Table 8. Pairwise comparison of ability parameter RMSE in 3 PL models by prior distribution type.

Prior distribution type	Difference	Standard error	<i>t</i>	<i>p</i>
Normal – Left skewed	-1.376		-0.858	0.911
Normal – Right skewed	-1.300		-0.811	0.926
Normal – Leptokurtic	-6.463		-4.030	0.001**
Normal – Platykurtic	-0.697		-0.434	0.992
Left skewed – Right skewed	0.076	1.604	0.047	0.999
Left skewed – Leptokurtic	-5.088		-3.172	0.017*
Left skewed – Platykurtic	0.679		0.424	0.993
Right skewed – Leptokurtic	-5.163		-3.219	0.015*
Right skewed – Platykurtic	0.604		-0.376	0.996
Leptokurtic – Platykurtic	5.767		3.595	0.004**

* $p < .05$, ** $p < .01$

Table 8 shows that the priori distribution type on the ability parameter RMSE in the 3 PL model is normal to leptokurtic ($t = -4.030$; $p < .01$), left skewed to leptokurtic ($t = -3.172$; $p < .05$), right skewed to leptokurtic ($t = -3.219$; $p < .05$), significant differences were found between leptokurtic and platykurtic ($t = 3.595$; $p < .01$). In the 3 PL model, RMSE increase as the priori distribution becomes leptokurtic on the ability parameters. The RMSE change according to the priori distribution type on the ability parameter in the 3 PL model is given in Figure 5.

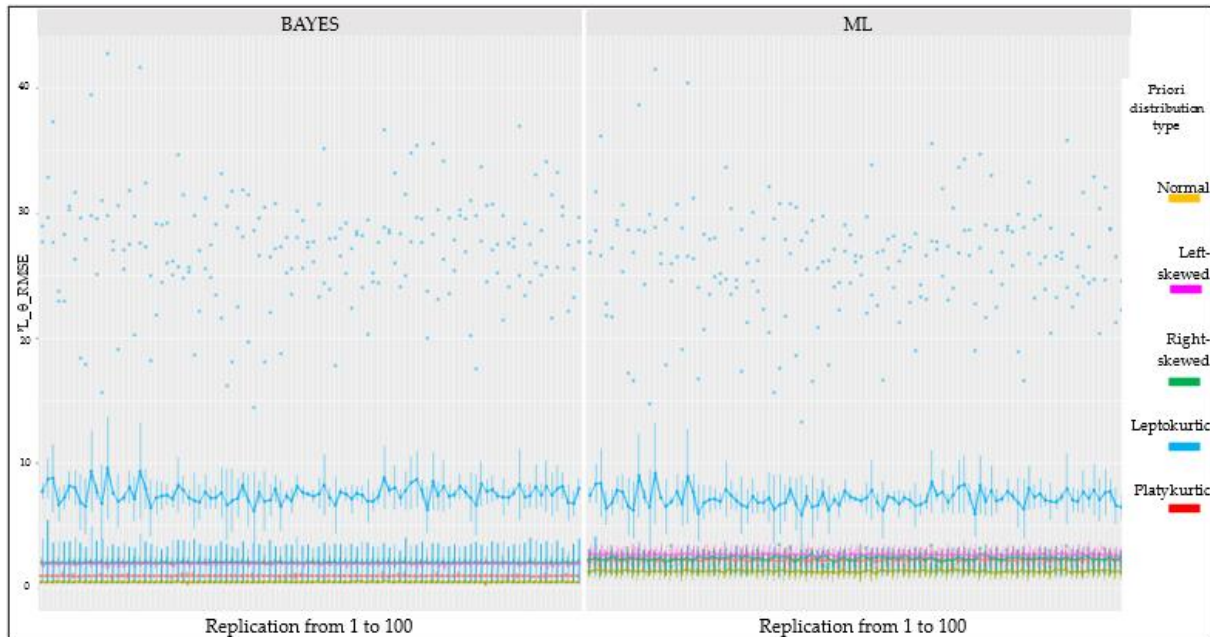
Figure 5. The change of ability parameter RMSE in the 3 PL model by prior distribution type.

Figure 5 shows that the priori distribution type becomes leptokurtic in the 3 PL model, and the RMSE of ability parameters takes higher values. However, according to the ability parameters estimation method, other distribution types did not differentiate on the priori distribution type, except for the leptokurtic distribution. Therefore, as in the 2 PL model, the leptokurtic priori distribution on the estimations of the ability parameters significantly affects the RMSE. This is seen in both ML and Bayesian estimation methods. Accordingly, the leptokurtic of the priori distribution harms the RMSE of the ability parameters, regardless of the model (2 PL or 3 PL). This situation is likely caused by the leptokurtic distribution (lower standard deviation and narrow ranges) and the data structure generated while performing the simulation. For this reason, cases where priori is leptokurtic should be examined in more detail within the framework of IRT parameter estimations.

3.3. Investigation of a_{RMSE} , b_{RMSE} , c_{RMSE} Estimated by ML and Bayesian Methods in 2 PL Model

RMSE changes of item parameters according to sample size, test length, and estimation method in 2 PL models with normal and non-normal (left skewed, right skewed, leptokurtic, and platykurtic) priori distribution stated in the third problem of the study were analyzed by mixed model ANOVA. Accordingly, the mixed model ANOVA results were performed for the item parameters according to sample size, test length, and estimation method in the data in 2 PL models with normal and non-normal priori distribution (left-skewed, right-skewed, leptokurtic, and platykurtic) are given in Table 9.

Table 9 shows that according to the mixed model ANOVA results for the item discrimination parameter RMSE in the data in the 2 PL models with normal and non-normal priori distribution, the main effects of independent variables as estimation method ($F_{(1, 88)} = 8.17$; $p < .01$, $\eta^2 = .045$), sample size ($F_{(2, 87)} = 8.97$; $p < .01$, $\eta^2 = .090$) and priori distribution type ($F_{(4, 85)} = 3.93$; $p < .01$, $\eta^2 = .083$) have significant effects. Test length ($F_{(2, 87)} = 0.10$; $p > .05$, $\eta^2 = .001$) did not have a significant effect. Among the independent variables found to be statistically significant, the estimation method has a small effect size, the sample size has a medium effect size, and the priori ability distribution has a medium effect size.

Table 9. Mixed model ANOVA results for item parameters RMSE in data in 2 PL models with normal and non-normal priori distribution.

Independent variables	Mean squares of error	Degrees of freedom	F	p	Generalized η^2
<i>Item discrimination (a_{RMSE})</i>					
Estimation method (K)	5385.60	1	8.17	0.005**	0.045
Sample size (S)	4935.31	2	8.97	0.001**	0.090
Test length (M)	5939.41	2	0.10	0.905	0.001
Prior distribution type (D)	5141.71	4	3.93	0.006**	0.083
K*S	3891.39	2	7.52	0.001**	0.070
K*M	5621.35	2	0.05	0.952	0.001
K*D	4210.31	4	3.34	0.014*	0.069
Error	53.68	198			
Total	35178				
<i>Item difficulty (b_{RMSE})</i>					
Estimation method (K)	5827.21	1	1.26	0.264	0.002
Sample size (S)	5706.21	2	2.08	0.131	0.007
Test length (M)	5900.27	2	0.58	0.562	0.002
Prior distribution type (D)	5606.80	4	1.94	0.111	0.014
K*S	5597.01	2	1.69	0.191	0.006
K*M	5931.07	2	0.65	0.523	0.003
K*D	5244.57	4	2.37	0.060	0.017
Error	311.54	198			
Total	40124.68				

* $p < .05$, ** $p < .01$

According to the mixed model ANOVA results, none of the independent variables created a significant difference for the item difficulty parameter RMSE values in the data in the 2 PL model with and without normal a priori ability distribution. Therefore, only significant conditions on the item discrimination parameter RMSE were given in the third research problem.

In the 2 PL model, sample size with estimation method ($F_{(2, 84)} = 7.52$; $p < .01$, $\eta^2 = .070$) and prior distribution type with estimation method ($F_{(4, 80)} = 3.34$; $p < .05$, $\eta^2 = .069$) were significant differences on item discrimination parameter RMSE. However, these pairwise interactions had moderate effect sizes. Therefore, for the data in the 2 PL model, the pairwise comparisons of the estimation method having a significant effect on the item discrimination parameter are given in Table 10.

Table 10. Pairwise comparisons of item discrimination parameter RMSE in 2 PL model by method of estimation.

Estimation method	Difference	Standard error	t	p
ML-Bayes	4.421	1.547	2.858	0.005**

* $p < .05$, ** $p < .01$

Table 10 shows that the estimation method on the item discrimination parameter RMSE in the 2 PL model data with normal and non-normal priori distribution type is in favor of the Bayesian estimation method and significant ($t=2.858$; $p < .01$). RMSE changes of the item discrimination parameter estimation methods in the 2 PL model are given in Figure 6.

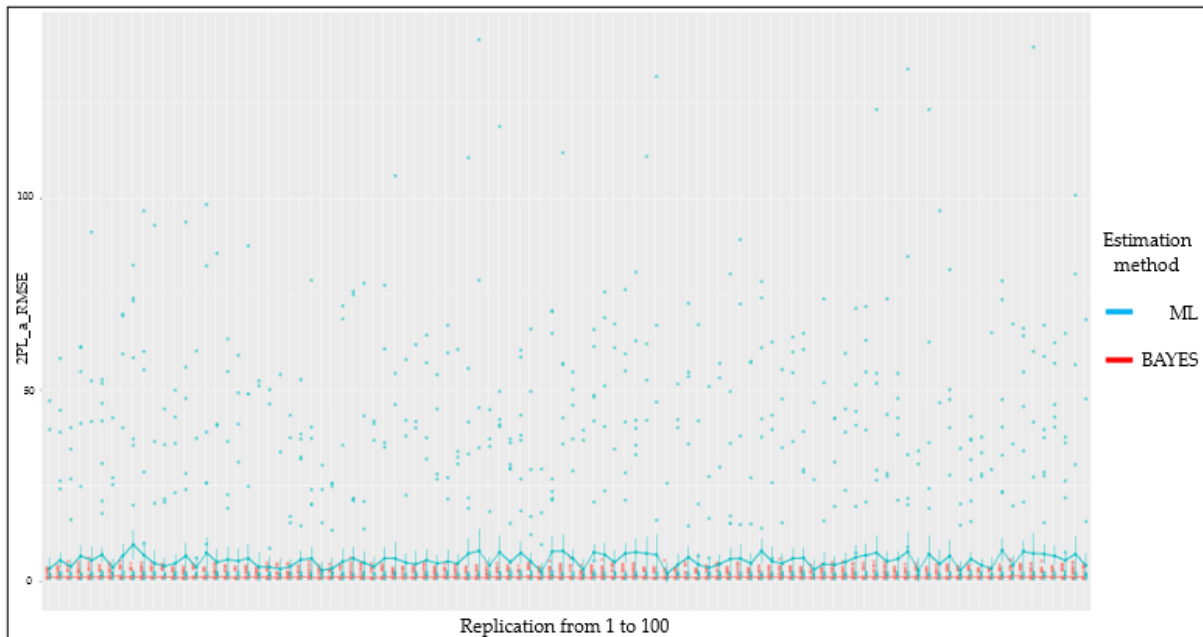
Figure 6. The change of item discrimination parameter RMSE in 2 PL model by method of estimation

Figure 6 shows that the item discrimination parameter RMSE in the 2 PL model, independent of all research conditions, takes lower Bayesian estimation values than ML estimation. Furthermore, while the item discrimination parameter RMSE (a_{RMSE}) shows a scattering according to the estimation results of the ML method, these values are more linear and stable in Bayesian estimation. Accordingly, the Bayesian approach provides advantages over the ML procedure in estimating item discrimination parameters. Pairwise comparisons of the sample size significantly affected the item discrimination parameters for the data in the 2 PL models, which are given in Table 11.

Table 11. Pairwise comparisons of item discrimination parameter RMSE in 2 PL model by sample size and estimation method.

Estimation method	Sample size	Difference	Standard error	t	p
ML	100 500	11.958	2.278	5.250	0.000**
	100 1000	12.164		5.340	0.000**
	500 1000	0.207		0.091	0.999
Bayes	100 500	1.198	2.278	0.526	0.995
	100 1000	1.290		0.566	0.993
	500 1000	0.092		0.040	0.999
ML*Bayes	100 100	-11.633	2.278	-5.107	0.000**
	500 500	-0.873		-0.383	0.999
	1000 1000	-0.758		-0.333	0.999

* $p < .05$, ** $p < .01$

Table 11 shows a significant difference between the RMSE of the item discrimination parameter estimated by the ML method in the 2 PL model between sample sizes of 100 and 500 ($t=5.250$; $p<.01$) and between 100 and 1000 ($t=5.340$; $p<.01$). However, there was no difference between sample sizes in Bayesian estimation. Accordingly, the significant RMSE in small samples in ML estimation decreased in the Bayesian method. Nevertheless, the RMSE of the item discrimination parameter estimated by different methods at the same sample sizes showed a significant difference at a sample size of 100 ($t=-5.107$; $p<.01$). This difference was eliminated as the sample size increased. Accordingly, using the Bayesian estimation method to obtain item discrimination parameters with low RMSE in small samples is more suitable. RMSE change

according to sample size on item discrimination parameter in the 2 PL model is given in Figure 7.

Figure 7. The change of item discrimination parameter RMSE in 2 PL model by sample size.

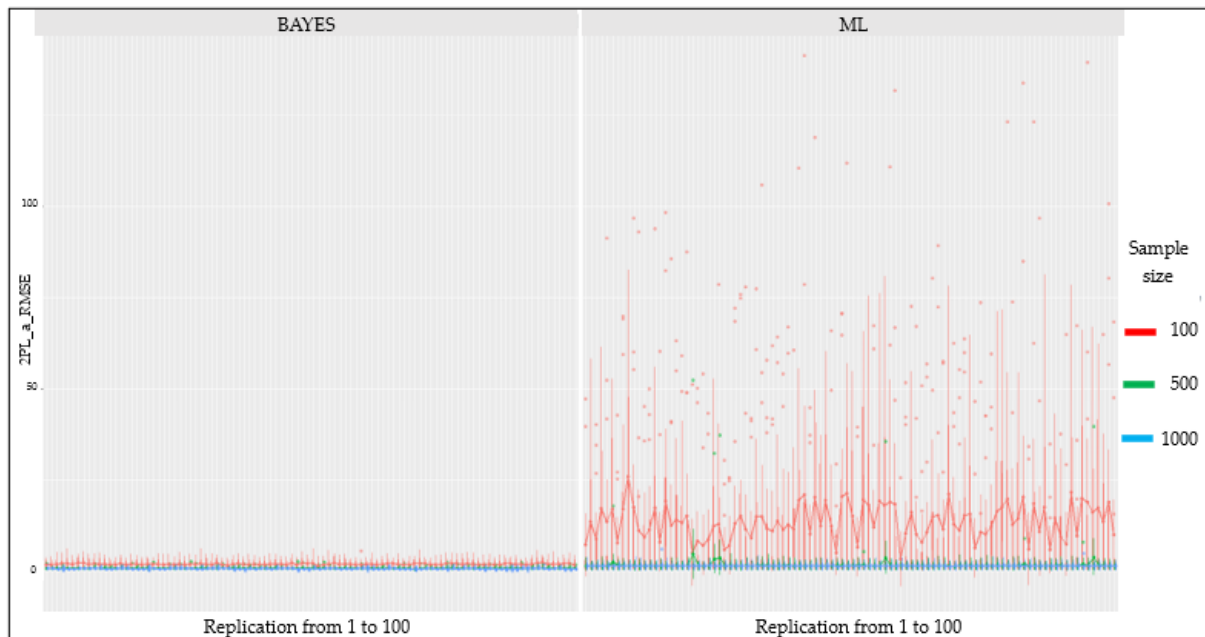


Figure 7 shows that the Bayesian estimation method produced lower values on item discrimination parameter RMSE (a_{RMSE}) when the sample size decreased compared to ML estimation. When the sample size decreased to 100 in the ML estimation, the item discrimination parameter RMSE increased excessively and created scattering. In this case, when the Bayesian estimation method was used, RMSE tended to decrease and showed a linear distribution. When the sample size was 500 or 1000, RMSE did not show a significant difference according to the estimation method. As can be understood from this, when the ML estimation method is used in the 2 PL model, a sample of at least 500 sample size should be used to reduce the item discrimination parameter RMSE. When the sample size drops to 100, the Bayesian estimation method should be used. Pairwise comparisons on the item discrimination parameter in the 2 PL model according to the priori distribution form are given in Table 12.

Table 12 shows that significant differences were found between the item discrimination parameter RMSE estimated by ML method in the 2 PL model between normal and left skewed ($t=-4.031$; $p<.01$), normal and right skewed ($t=-3.754$; $p<.05$), left skewed and leptokurtic ($t=3.815$; $p<.01$), left skewed and platykurtic ($t=3.513$; $p<.05$), right skewed and leptokurtic ($t=3.538$; $p<.05$) and right skewed and platykurtic ($t=3.236$; $p<.05$) according to the distribution types. These differences were eliminated in Bayesian estimation. The RMSE of the item discrimination parameter estimated by Bayesian method in the 2 PL model were not significantly affected by the type of prior distribution. In the same type of a priori distributions, item discrimination parameter RMSE estimated by ML and Bayesian methods differed significantly when the distribution was left skewed ($t=3.569$; $p<.05$) or right skewed ($t=3.300$; $p<.05$). However, according to the estimation methods, no difference was found for the other distribution types. In the 2 PL model, RMSE on the item discrimination parameter according to the priori ability distribution types are given in Figure 8.

Table 12. Pairwise comparisons of item discrimination parameter RMSE in 2 PL model by priori distribution type and estimation method.

Estimation method	Prior distribution type	Difference	Standard error	<i>t</i>	<i>p</i>		
ML	Normal	Left skewed	-12.330		-4.031	0.004**	
		Right skewed	-11.482		-3.754	0.011*	
		Leptokurtic	-0.659		-0.216	0.999	
		Platykurtic	-1.584		-0.518	0.999	
		Right skewed	0.847		0.277	0.999	
	Left skewed	Leptokurtic	11.670		3.815	0.009**	
		Platykurtic	10.746		3.513	0.024*	
		Right skewed	Leptokurtic	10.823		3.538	0.022*
			Platykurtic	9.898		3.236	0.050*
		Leptokurtic	Platykurtic	-0.925		-0.302	0.999
Bayes	Normal	Left skewed	-1.162		-0.380	0.999	
		Right skewed	-1.137		-0.372	0.999	
		Leptokurtic	-0.329	3.059	-0.108	0.999	
		Platykurtic	-0.061		0.020	0.999	
		Right skewed	0.025		0.008	0.999	
	Left skewed	Leptokurtic	0.833		0.272	0.999	
		Platykurtic	1.101		0.360	0.999	
		Right skewed	Leptokurtic	0.808		0.264	0.999
			Platykurtic	1.076		0.352	0.999
		Leptokurtic	Platykurtic	0.268		0.088	0.999
ML*Bayes	Normal	Normal	-0.252		-0.082	0.999	
	Left skewed	Left skewed	10.916		3.569	0.020*	
	Right skewed	Right skewed	10.093		3.300	0.044*	
	Leptokurtic	Leptokurtic	0.078		0.026	0.999	
	Platykurtic	Platykurtic	1.271		0.416	0.999	

p* < .05, *p* < .01

Figure 8. The change of item discrimination parameter RMSE in 2 PL model by priori distribution type.

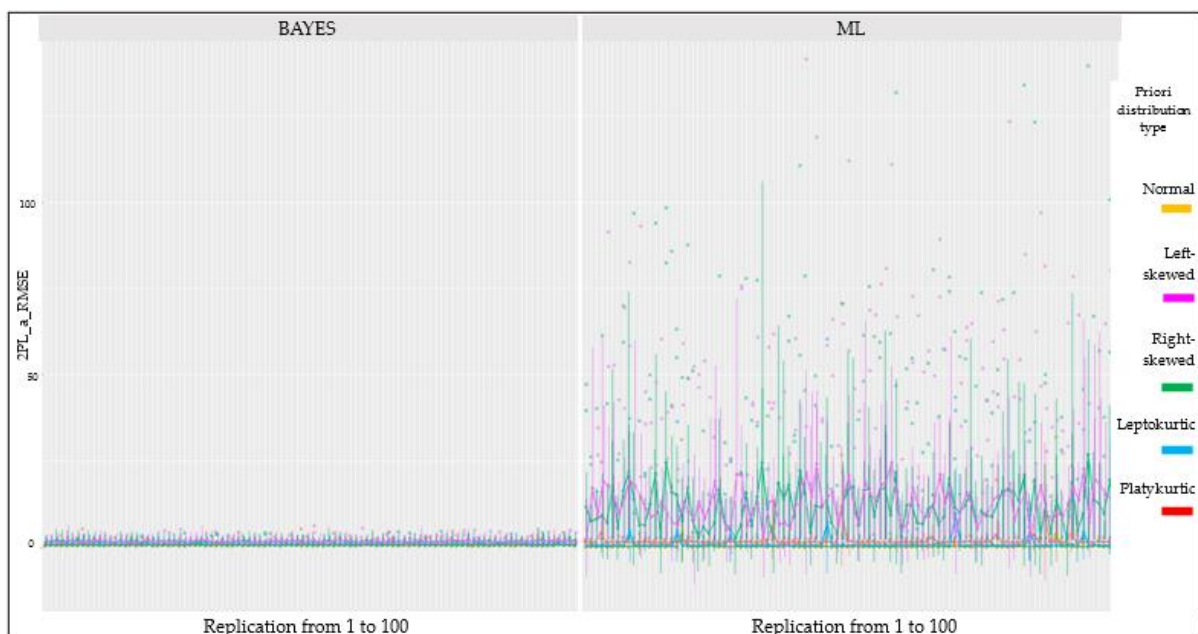


Figure 8 shows that the item discrimination parameter RMSE (a_{RMSE}) is higher as the priori distribution becomes skewed in the 2 PL model. In the ML estimation method, the item discrimination parameter RMSE (a_{RMSE}) increases as the priori distribution becomes skewed to the left or right. The leptokurtic or platykurtic of the prior distribution does not have an increasing effect on the item discrimination parameter RMSE. However, the Bayesian estimation method reduced the high RMSE of the item discrimination parameter if the priori ability distribution was skewed to the left or right. Accordingly, in the 2 PL model, the item discrimination parameter RMSE (a_{RMSE}) is affected by the differentiation of the priori distribution type. As a result, it shows low values when using the Bayesian estimation method.

3.4. Investigation of a_{RMSE} , b_{RMSE} , c_{RMSE} Estimated by ML and Bayesian Methods in 3 PL Model

RMSE changes of item parameters according to sample size, test length, and estimation method in 3 PL models with normal and non-normal (left skewed, right skewed, leptokurtic, and platykurtic) priori distribution stated in the fourth problem of the research were analyzed by mixed model ANOVA. Accordingly, the mixed model ANOVA results were performed for the item parameters according to sample size, test length, and estimation method in the data in 3 PL models with normal and non-normal priori distribution (left skewed, right skewed, leptokurtic, and platykurtic) are given in Table 13.

Table 13 shows that according to the mixed model ANOVA results for the item discrimination parameter RMSE in the data in the 3 PL models with normal and non-normal priori distribution, the main effects of independent variables as estimation method ($F_{(1, 88)} = 28.61$; $p < .01$, $\eta^2 = .203$), sample size ($F_{(2, 87)} = 4.55$; $p < .05$, $\eta^2 = .078$) and priori distribution type ($F_{(4, 85)} = 6.40$; $p < .01$, $\eta^2 = .192$) had significant effects. Test length ($F_{(2, 87)} = 0.53$; $p > .05$, $\eta^2 = .010$) did not show a significant difference. Among the independent variables found to be statistically significant, the estimation method is high, the sample size is medium, and the priori ability distribution type has a high effect size. In the 3 PL model, sample size ($F_{(2, 84)} = 5.22$; $p < .01$, $\eta^2 = .085$) has a significant and moderate effect size and priori distribution type ($F_{(4, 80)} = 13.46$; $p < .01$, $\eta^2 = .295$) has a significant and high effect size on item discrimination parameter RMSE. According to the mixed model ANOVA results for the item difficulty parameter RMSE in the 3 PL models with normal and non-normal priori distribution, none of the independent variables created a significant difference.

According to the mixed model ANOVA results for lower asymptote parameter RMSE in the data in 3 PL models with normal and non-normal priori distribution, estimation method ($F_{(1, 88)} = 9.10$; $p < .01$, $\eta^2 = .074$) and priori distribution type ($F_{(4, 80)} = 13.00$; $p < .01$, $\eta^2 = .306$) as the main effects of independent variables created significant differences. Sample size ($F_{(2, 87)} = 2.49$; $p > .05$, $\eta^2 = .043$) and test length ($F_{(2, 87)} = 0.50$; $p > .05$, $\eta^2 = .009$) were not significantly different. The estimation method that created a significant difference had a medium effect size, and the priori distribution type had a high effect size. In the 3 PL model, the estimation method from interactions and priori distribution type ($F_{(4, 80)} = 4.11$; $p < .01$, $\eta^2 = .117$) had a significant and medium effect size on lower asymptote parameter RMSE. Pairwise comparisons of the estimation method's significant difference in the item discrimination parameter for the data in the 3 PL model are given in Table 14.

Table 13. Mixed model ANOVA results for item parameters RMSE in 3 PL models with normal and non-normal priori distribution.

Independent variables	Mean squares of error	Degrees of freedom	F	p	Generalized η^2
<i>Item discrimination (a_{RMSE})</i>					
Estimation method (K)	98140.76	1	28.61	0.001**	0.203
Sample size (S)	119088.76	2	4.55	0.013*	0.078
Test length (M)	129940.17	2	0.53	0.588	0.010
Prior distribution type (D)	103482.30	4	6.40	0.001**	0.192
K*S	79972.48	2	5.22	0.007**	0.085
K*M	99643.32	2	0.64	0.530	0.012
K*D	44745.66	4	13.46	0.001**	0.295
Error	273.87	198			
Total	675287.32				
<i>Item difficulty (b_{RMSE})</i>					
Estimation method (K)	1149170.54	1	1.82	0.180	0.001
Sample size (S)	1132417.18	2	2.08	0.132	0.001
Test length (M)	117180079	2	0.54	0.582	0.001
Prior distribution type (D)	1106887.16	4	2.06	0.093	0.001
K*S	1095077.60	2	2.03	0.138	0.001
K*M	1173689.34	2	0.54	0.586	0.001
K*D	1037641.31	4	2.16	0.081	0.001
Error	797927.50	198			
Total	8664611.42				
<i>Lower asymptote (c_{RMSE})</i>					
Estimation method (K)	0.06	1	9.10	0.003**	0.074
Sample size (S)	0.06	2	2.49	0.089	0.043
Test length (M)	0.07	2	0.50	0.606	0.009
Prior distribution type (D)	0.04	4	13.00	0.001**	0.306
K*S	0.06	2	2.11	0.127	0.037
K*M	0.06	2	0.32	0.727	0.006
K*D	0.03	4	4.11	0.004**	0.117
Error	0.00	198			
Total	0.38				

* $p < .05$, ** $p < .01$ **Table 14.** Pairwise comparisons of item discrimination parameter RMSE in 3 PL model by method of estimation.

Estimation method	Difference	Standard error	t	p
Bayes-ML	-35.323	6.604	-5.348	0.001**

* $p < .05$, ** $p < .01$

Table 14 shows that the item discrimination parameter RMSE of the data in the 3 PL models with normal and non-normal priori distribution were significant in favor of the Bayesian estimation method ($t = -5.348$; $p < .01$). Bayesian estimation method produced lower RMSE than the ML estimation method. RMSE changes of the item discrimination parameter (a_{RMSE}) estimation methods in the 2 PL model are given in Figure 9.

Figure 9. The change of item discrimination parameter RMSE values in 3 PL model by estimation methods.

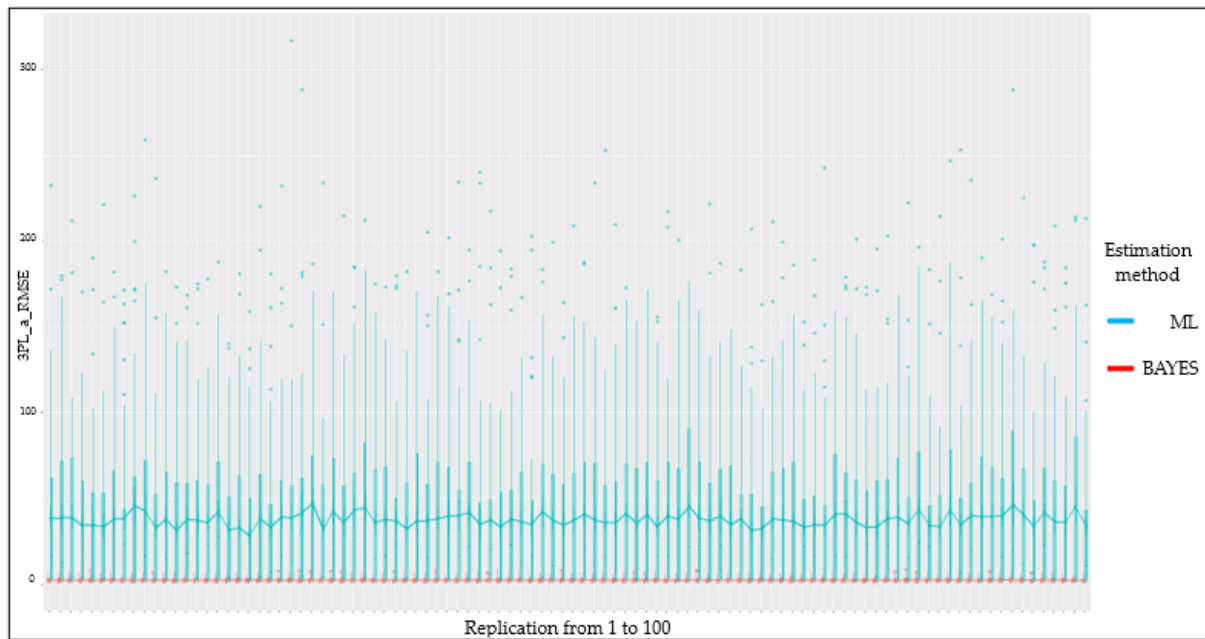


Figure 9 shows that the item discrimination parameter RMSE (a_{RMSE}) in the 3 PL model, independent of all simulation conditions, takes lower Bayesian estimation values than ML estimation. For the data in the 3 PL model, the pairwise comparisons of the sample size having a significant effect on the item discrimination parameter RMSE are given in Table 15.

Table 15. Pairwise comparisons of item discrimination parameter RMSE values by sample size and estimation method in 3 PL model.

Estimation method	Sample size	Difference	Standard error	t	p
ML	100 500	40.622	10.326	3.934	0.002**
	100 1000	46.255		4.479	0.001**
	500 1000	5.633		0.546	0.994
Bayes	100 500	2.711		0.263	0.999
	100 1000	2.941		0.285	0.999
	500 1000	0.231		0.022	0.999
ML*Bayes	100 100	62.398	6.043	0.001**	
	500 500	24.487	2.371	0.178	
	1000 1000	19.085	1.848	0.441	

* $p < .05$, ** $p < .01$

Table 15 shows that there is a significant difference between the item discrimination parameter RMSE estimated by ML method in the 3 PL model between sample sizes 100 and 500 ($t=3.934$; $p < .01$) and 100 and 1000 ($t=4.479$; $p < .01$), but no significant difference between 500 and 1000 ($t=0.546$; $p > .05$). However, using Bayes as the estimation method eliminated the significant differences between the sample sizes. Accordingly, using the Bayesian estimation method to estimate the item discrimination parameter more accurately in 3 PL models and small samples is more appropriate. Supporting this, when the sample size was 100 ($t=6.043$; $p < .01$), a significant difference was found between the RMSE of the item discrimination parameter according to the ML and Bayesian estimation method analyses. However, this significant difference is not observed as the sample size increases. RMSE change according to sample size on item discrimination parameter in the 3 PL model is given in Figure 10.

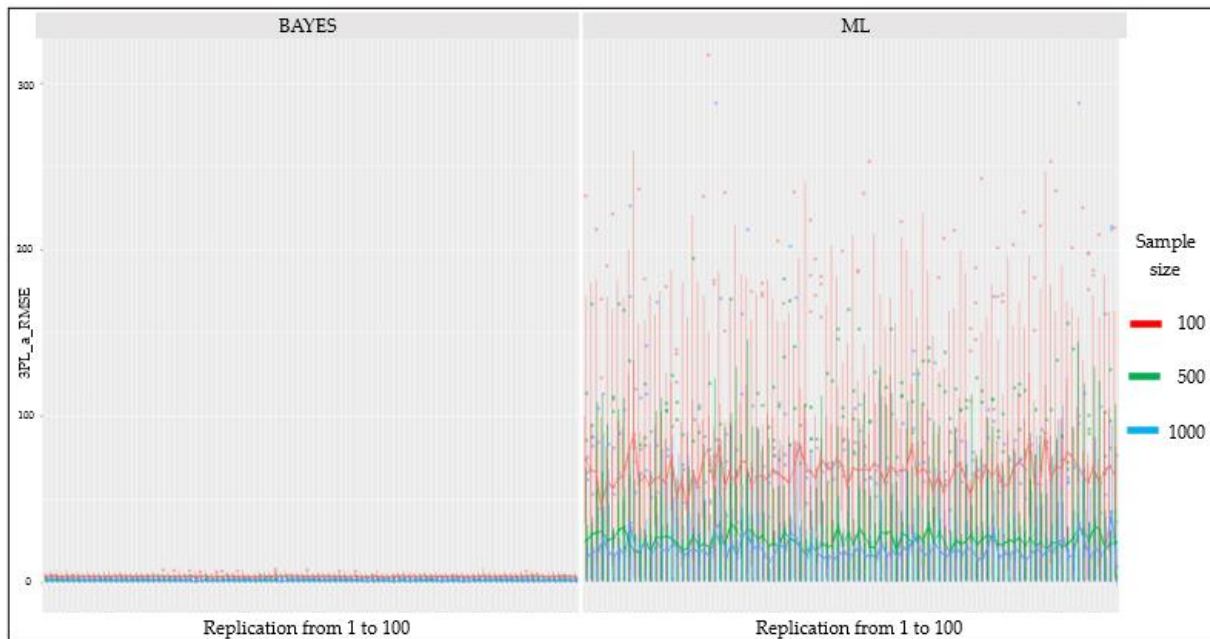
Figure 10. The change of item discrimination parameter RMSE by sample size in the 3 PL model.

Figure 10 shows that when the sample size decreased, the Bayesian estimation method produced lower RMSE on item discrimination parameters than ML estimation. In ML estimation, item discrimination RMSE increases as the sample size decreases. In addition, these values show scattering. This situation is similar to the results obtained in the 2 PL model. These values decrease as the sample size increases. However, the Bayesian method tends to reduce the item discrimination parameter RMSE compared to the ML method. In Bayesian estimation, the increase in sample size did not make a difference in the item discrimination parameter RMSE (a_{RMSE}). RMSE obtained according to sample size is linear. In other words, the Bayesian method reduced and stabilized the item discrimination parameter RMSE (a_{RMSE}) compared to the ML estimation. Pairwise comparisons on the item discrimination parameter in the 3 PL model according to the priori distribution type are given in Table 16.

Table 16 shows that significant differences were found between the RMSE of the item discrimination parameter estimated by the ML method in the 3 PL model between normal and right-skewed ($t=-8.852$; $p<.01$), normal and leptokurtic ($t=-4.516$; $p<.01$), left-skewed and right-skewed ($t=-7.960$; $p<.01$), left-skewed and leptokurtic ($t=-3.624$; $p<.05$), right-skewed and leptokurtic ($t=4.337$; $p<.01$), right-skewed and platykurtic ($t=8.400$; $p<.01$), leptokurtic and platykurtic ($t=4.063$; $p<.01$) according to the distribution types. However, no significant difference was found between the a priori distribution types when the same parameter was estimated with the Bayesian method. Bayesian estimation method eliminated the significant difference depending on the a priori distribution type. Confirming this, the item discrimination parameter RMSE estimated by ML and Bayesian methods in the same a priori distribution types show a significant difference when the distribution is right skewed ($t=9.274$; $p<.01$) or leptokurtic ($t=5.162$; $p<.01$). Here, unlike in the 2 PL model, a distribution of a priori leptokurtic in the 3 PL model was found to cause differentiation. No differentiation was observed for the other distribution types. RMSE on item discrimination parameters in the 3 PL model according to priori distribution type is given in Figure 11.

Table 16. Pairwise comparisons of item discrimination parameter RMSE in 3 PL model by priori distribution type and estimation method.

Estimation method	Prior distribution type	Difference	Standard error	<i>t</i>	<i>p</i>		
ML	Normal	Left skewed	-8.894		-0.892	0.996	
		Right skewed	-88.270		-8.852	0.001**	
		Leptokurtic	-45.027		-4.516	0.001**	
		Platykurtic	-4.508		0.452	0.999	
	Left skewed	Right skewed	-79.377		-7.960	0.001**	
		Leptokurtic	-36.134		-3.624	0.016*	
		Platykurtic	4.386		0.440	0.999	
		Leptokurtic	43.243		4.337	0.001**	
	Right skewed	Leptokurtic	83.763		8.400	0.001**	
		Platykurtic	40.520		4.063	0.004**	
		Leptokurtic	Platykurtic				
			Left skewed	-0.536		-0.054	0.999
Normal	Right skewed		-2.614	9.972	-0.262	0.999	
	Leptokurtic		-0.374		-0.038	0.999	
	Platykurtic	-0.653		-0.065	0.999		
	Right skewed	-2.077		-0.208	0.999		
Bayes	Left skewed	Leptokurtic	0.162		0.016	0.999	
		Platykurtic	-0.117		-0.012	0.999	
		Leptokurtic	2.239		0.225	0.999	
		Platykurtic	1.961		0.197	0.999	
Right skewed	Leptokurtic	-0.279		-0.028	0.999		
	Platykurtic						
	Normal	Normal	6.819		0.684	0.999	
	Left skewed	Left skewed	15.176		1.522	0.879	
ML*Bayes	Right skewed	Right skewed	92.476		9.274	0.001**	
	Leptokurtic	Leptokurtic	51.472		5.162	0.001**	
	Platykurtic	Platykurtic	10.673		1.070	0.986	

* $p < .05$, ** $p < .01$

Figure 11. The change of item discrimination parameter RMSE in the 3 PL model by priori distribution types.

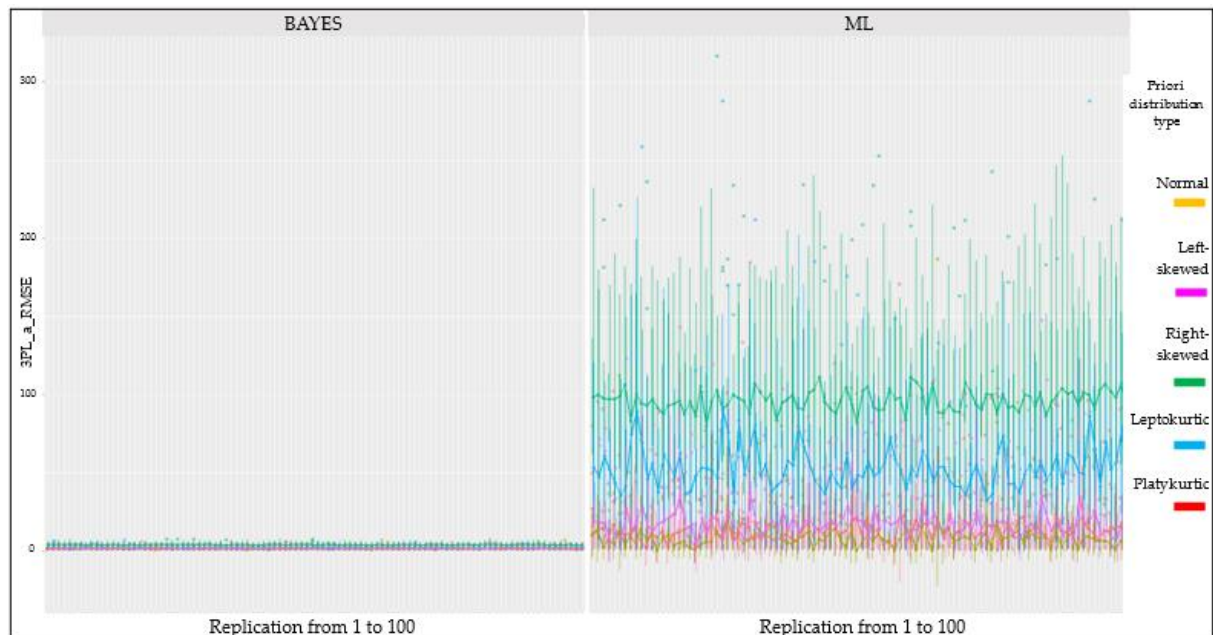


Figure 11 shows that the priori distribution becomes skewed in the 3 PL model, and the item discrimination parameter RMSE takes higher values. These high RMSE were reduced by the Bayesian estimation method. In the ML estimation in the 3 PL model, the item discrimination parameter RMSE gave the highest results when the priori distribution was skewed to the right. This was followed by leptokurtic, left skewed, platykurtic, and normal distributions. The fact that the model is 3 PL is an essential factor for the item discrimination parameter RMSE (a_{RMSE}) to be the highest when the priori distribution is skewed to the right. Unlike the 2 PL model, by adding a third parameter, the lower asymptote parameter (c_i) in this model changes the starting point of the priori distributions. Therefore, the most affected by this situation are the right-skewed priori parameters. When Bayesian estimation was used, the item discrimination parameter RMSE (a_{RMSE}) produced lower RMSE in all a priori distributions compared to ML estimation, which was stably distributed. Pairwise comparisons of the estimation method's significant effect on the lower asymptote parameter for the data in the 3 PL model are given in Table 17.

Table 17. Pairwise comparisons of lower asymptote parameter RMSE in 3 PL model by estimation method.

Estimation method	Difference	Standard error	t	p
Bayes-ML	0.016	0.005	3.016	0.003**

* $p < .05$, ** $p < .01$

Table 17 shows that the lower asymptote parameter is significant and in favor of the Bayesian estimation method on RMSE in 3 PL models with normal and non-normal priori distribution ($t=3.016$; $p < .01$). Bayesian estimation method produced lower RMSE than the ML estimation method. RMSE changes of the lower asymptote parameter estimation methods in the 3 PL model are given in Figure 12.

Figure 12. The change of the lower asymptote parameter RMSE in the 3 PL model by estimation methods.

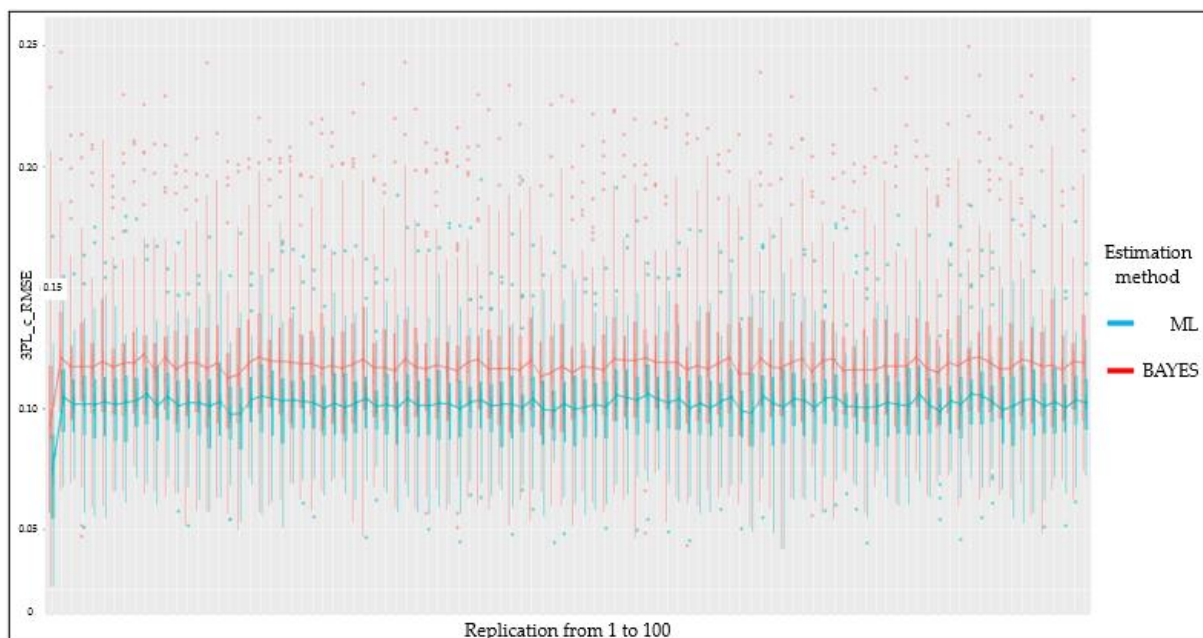


Figure 12 shows that the RMSE of the lower asymptote parameter in the 3 PL model takes higher values in Bayesian estimation regardless of the research conditions. Unlike other parameters, ML estimation was more effective than Bayesian estimation in decreasing the RMSE of the lower asymptote parameters. There are few studies on the lower asymptote parameter in the literature. This result is likely due to the distribution type defined for the lower asymptote parameter while creating the function for the priori distribution. The data in the 3 PL

model's pairwise comparisons of the priori distribution type that significantly affect the lower asymptote parameter is given in Table 18.

Table 18. Pairwise comparisons of the lower asymptote parameter RMSE in the 3 PL model by priori distribution type and estimation method.

Estimation method	Prior distribution type	Difference	Standard error	<i>t</i>	<i>p</i>		
ML	Normal	Left skewed	0.017	2.012	0.593		
		Right skewed	-0.007	-0.765	0.998		
		Leptokurtic	-0.015	-1.742	0.768		
		Platykurtic	-0.003	-0.333	0.999		
	Left skewed	Right skewed	-0.024	-2.777	0.163		
		Leptokurtic	-0.032	-3.754	0.011*		
		Platykurtic	-0.020	-2.346	0.372		
		Leptokurtic	-0.008	-0.977	0.993		
	Right skewed	Platykurtic	0.004	0.431	0.999		
		Leptokurtic	0.012	1.409	0.921		
		Leptokurtic	Platykurtic	0.043	5.102	0.001**	
			Right skewed	0.024	2.816	0.149	
Leptokurtic	-0.023		0.009	-2.693	0.194		
Platykurtic	0.020		2.402	0.339			
Bayes	Left skewed	Right skewed	-0.019	-2.286	0.509		
		Leptokurtic	-0.066	-7.795	0.001**		
		Platykurtic	0.043	5.094	0.001**		
		Leptokurtic	-0.047	-5.509	0.001**		
	Right skewed	Platykurtic	-0.004	-0.414	0.999		
		Leptokurtic	0.043	5.094	0.001**		
		ML*Bayes	Normal	Normal	-0.030	-3.544	0.022*
			Left skewed	Left skewed	-0.004	-0.454	0.999
Right skewed	Right skewed		0.000	0.037	0.999		
Leptokurtic	Leptokurtic		-0.038	-4.494	0.001**		
Platykurtic	Platykurtic	-0.007	-0.809	0.998			

* $p < .05$, ** $p < .01$

Table 18 shows a significant difference between the lower asymptote parameter RMSE values estimated by ML method in 3 PL models between left skewed and leptokurtic ($t = -3.754$; $p < .05$) according to distribution types. In Bayesian estimation, there is a significant difference between normal and left skewed ($t = 5.102$; $p < .01$), left skewed and leptokurtic ($t = -7.795$; $p < .01$), left skewed and platykurtic ($t = 5.094$; $p < .01$), right skewed and leptokurtic ($t = -5.509$; $p < .01$), leptokurtic and platykurtic ($t = 5.094$; $p < .01$) according to distribution types. As with the other parameters, no significance is expected for this parameter. However, the advantages of Bayesian estimation over ML estimation were not observed at lower asymptote parameters. The lower asymptote parameter RMSE estimated by ML and Bayesian methods in the same priori distribution types showed a significant difference in the normal ($t = -3.544$; $p < .05$) and leptokurtic ($t = -4.494$; $p < .01$) distributions. No difference was observed in other distribution types. Lower asymptote parameter RMSE according to the priori ability distribution type in the 3 PL model are given in Figure 13.

Figure 13. The change of the lower asymptote parameter RMSE in the 3 PL model by priori distribution types.

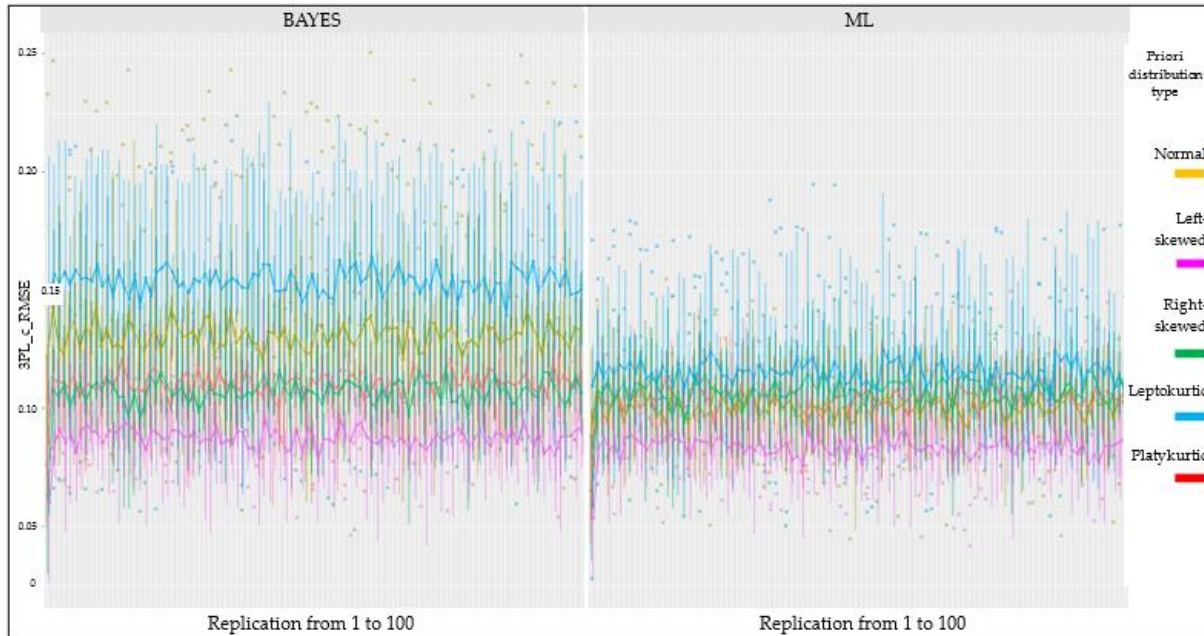


Figure 13 shows that the priori distribution becomes skewed in the 3 PL model; the lower asymptote parameter RMSE takes higher values. However, as the type of priori distribution becomes leptokurtic, the lower asymptote parameter RMSE increases in Bayesian estimation, unlike the other item parameters. Accordingly, the ML estimation method produced lower RMSE as the priori distribution became leptokurtic in the 3 PL model. In addition, the RMSE obtained in the ML estimation for all priori distribution types was distributed in a narrower area than Bayesian estimation. The lower asymptote parameter RMSE (CR_{MSE}) obtained from Bayesian estimation is spread over a wider area because the initial parameter values are generated with a distribution other than the normal distribution. Standard Bayesian estimations tend to normalize the posterior distribution because the priori distribution is normal.

4. DISCUSSION and CONCLUSION

Considering the conditions in all the problems of the research, in the first research problem in which the RMSE of the ability parameters were examined, In the data in the 2 PL model, the estimation method on the RMSE of the ability parameters, test length, the type of priori distribution, and the interaction between estimation method and the priori distribution type created significant differentiation. These results are like the results of Finch and Edwards (2015) when examined in general terms. Likewise, Bayesian estimations give more accurate results in cases where the latent feature is non-normally distributed in the 2 PL model. A similar situation in terms of test length is also seen in Köse (2010)'s study. The change in test length affects the estimation results in ability parameters. An increase in test length decreases the RMSE of ability parameters.

In the second research problem, test length and priori distribution type created significant differences in the RMSE of the ability parameters in the data in the 3 PL model. The general results for this problem are like the results of Swaminathan and Gifford (1986). They suggested that their study use Bayesian estimation instead of ML for the 3 PL model. In addition, Karadavut (2019) stated in her research that when estimating the ability parameter in the 3 PL model, not knowing the priori distribution type would lead to erroneous estimations. A similar situation can be seen in this study's differentiation of the priori distribution type.

In estimating ability parameters and RMSE, the estimation method made a significant difference only in 2 PL models. This significance is in favor of the Bayesian estimation method.

Because Bayesian estimation reduced the high error values obtained in ML to lower values. Similar results were obtained in studies in the literature (Swaminathan & Gifford, 1986; Harwell & Janosky, 1991; Gao & Chen, 2005; Finch & Edwards, 2015).

In the third research problem, in which item parameters RMSE were examined, estimation method, sample size, priori distribution type, the interaction of estimation method and sample size, and interaction of estimation method and priori distribution type on item discrimination parameter RMSE in the data in the 2 PL model created significant differences. In the 2 PL model data, no condition caused a significant difference in the RMSE of the item difficulty parameter. These results are like Harwell and Janosky's (1991) results. Accordingly, Bayesian estimation is considered sufficient for small samples and short tests in the 2 PL model. It is stated in Stone's (1992) study that as the priori distribution for the item discrimination parameter becomes skewed, the bias in the ML estimation increases. In this study, the RMSE for the item discrimination parameter is also affected by the skewness of the prior distribution type. In this respect, these two studies showed similar results. It is also seen in the study of Sass et al. (2008) that item parameters are affected by priori distribution and produce high error values.

In the fourth research problem, the estimation method, sample size, priori distribution type, estimation method and sample size interaction, and estimation method and priori distribution type interaction on the item discrimination parameter RMSE in the data in the 3 PL model created significant differences. In the 3 PL model data, no conditions were significant on the item difficulty parameter RMSE. However, in the 3 PL model data, the estimation method on the RMSE of the lower asymptote parameter, the priori distribution type, and the interaction of the estimation method and the priori distribution type created significant differences. When these results are examined, it is seen that the suggestion of Swaminathan and Gifford (1986) is correct. Accordingly, this related research proposes the Bayesian method for parameter estimation for the 3 PL model. In this study, using the Bayesian estimation method in estimating item parameters in the 3 PL model, especially in the item discrimination parameter, provides an advantage. Likewise, as in the study of Gao and Chen (2005), Bayesian estimation gave more precise results in estimating item parameters when the sample size decreased to 100.

In estimating item parameters and RMSE, the estimation method generally showed a significant differentiation. This differentiation is significant for item discrimination RMSE (a_{RMSE}) and lower asymptote RMSE (c_{RMSE}) parameters regardless of the model. Bayesian estimation method for this significant differentiation item discrimination parameter; for the lower asymptote parameter, the ML estimation method is in favor. However, according to the estimation method for the item difficulty RMSE (b_{RMSE}) parameter, there is no differentiation between 2 PL and 3 PL models. This situation in the item difficulty parameter yielded similar results to the study of Kıbrıslıoğlu Uysal (2020).

While the sample size did not make a significant difference in estimating the ability parameter, the test length, the priori distribution type, and the estimation method (only in the 2 PL model) created significant differences in the RMSE. The sample size does not affect the ability of parameter estimation and error values because the number of estimated parameters is only one. This is similar to the research of Goldman and Raju (1986) and Harwell and Janosky (1991). The study of Goldman and Raju (1986) stated that the sample size of 250 would be sufficient when the estimated parameters were reduced to 1. Harwell and Janosky (1991) concluded that samples of 15 items and 250 people were sufficient. A similar situation is seen in the study of Şahin and Anıl (2017). Şahin and Anıl (2017) concluded that a sample of 150 people would be sufficient to make parameter estimation in 1 PL model.

The sample size was only effective in the RMSE estimations of the item discrimination parameter. This applies when both the 2 PL and 3 PL models are used. The increase in sample size positively affected the item discrimination parameter RMSE (a_{RMSE}), and these values decreased. However, as the sample size decreased, especially the RMSE of the item discrimination parameter showed excessive swelling. The swelling in the RMSE of the item

discrimination parameter (a_{RMSE}) due to estimation with the ML method was also seen in the studies of Chuah et al. (2006). However, the Bayesian estimation method played an important role in reducing this swelling. A similar situation is seen in the study of Gao and Chen (2005). In this study, it has been stated that Bayesian estimations give more accurate results than marginal maximum likelihood estimations when the sample size drops to 100.

Increasing the test length only decreased the ability parameter RMSE (Θ_{RMSE}). Moreover, in some cases where the test length is 40, the results of the ML and Bayesian methods for estimating ability have taken values close to each other. Similarly, Gao and Chen (2005) emphasized in their study that increasing test length and sample size tends to reduce the standard errors of estimations. However, when the test length decreased to 10, it caused swelling in the RMSE of the ability parameters in the ML estimation. However, this situation was reduced by the Bayesian estimation method. Item discrimination (a_i), item difficulty (b_i), and lower asymptote (c_i) parameters RMSE were not affected in any way by the test length change.

The priori distribution type ability parameters have significant differences in RMSE. According to the logistic model, the priori distribution type did not significantly differ in ability parameters. In both 2 PL and 3 PL models, the priori distribution type, item discrimination (a_i), and lower asymptote (c_i) parameters showed a significant difference in RMSE. In 2 PL and 3 PL models, there was no significant difference in item difficulty parameter RMSE (b_{RMSE}) values according to the priori distribution type. Differentiation of item parameters according to priori distribution type is more significant on the left and right skewed distributions than other distribution types. In a similar study conducted by Doğan (2002), distribution types (skewed or leptokurtic and platykurtic) affected the parameter invariance of the IRT. It was stated that the differentiation was higher in skewed distributions. A similar situation is observed in the studies of Seong (1990), Stone (1992), Kirisci et al. (2001), Sass et al. (2005) and Karadavut (2019).

The logistic model was significant on the RMSE of ability and item parameters. The 3 PL model produced higher prediction RMSE than the 2 PL model. The Bayesian estimation method decreased these values more than the ML.

The parameter estimation method, ability, and item parameters created a significant difference in the RMSE in different conditions that constitute the research's aim. In addition, it was shown that the Bayesian estimation method obtained lower RMSE than the ML estimation method in all simulation conditions. However, the significance of these RMSEs was observed in only some simulation conditions.

RMSE is the total error indicator of parameter estimation's precision and estimation bias (Thissen & Wainer, 1983). When the literature was reviewed, the standard errors of parameter estimation for commonly used models (Rasch, 1 PL, 2 PL, and 3 PL) needed to be comprehensively addressed (Lord, 1980). As stated in the study results, the Bayesian method reduced the RMSE of ability and item parameters to lower levels than the ML method.

Accordingly, the Bayesian estimation method seems advantageous since it produces lower parameter RMSE than the ML estimation method. Moreover, especially when the ML estimation method is used, it is seen that it tends to reduce the excessive increase in parameter RMSE that occurs in small samples and short tests.

Nowadays, it is possible to use IRT to develop classroom achievement tests. However, the first issue is how to do this with small samples and short tests. The Bayesian approach makes this possible and reduces the estimation errors to acceptable levels. In addition, it is only sometimes possible for the distribution under study to be normal. The ML estimation method does not give accurate results in such a case. At this point, the advantages of Bayesian estimation are utilized. The results of this study show that Bayesian estimation can be offered as a solution where ML estimation cannot obtain accurate results.

Acknowledgments

This research article was produced from the doctoral dissertation of first author under the supervision of second author.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors. **Ethics Committee Number:** Ankara University, 04.11.2019, 13-339.

Contribution of Authors

Eray Selçuk: Investigation, Resources, Visualization, Software, Coding, Analysis, and Writing-original draft. **Ergül Demir:** Investigation, Methodology, Supervision, Critical Review and Validation.

Orcid

Eray Selçuk  <https://orcid.org/0000-0003-4033-4219>

Ergül Demir  <https://orcid.org/0000-0002-3708-8013>

REFERENCES

- Akour, M., & Al-Omari, H. (2013). Empirical investigation of the stability of IRT item-parameters estimation. *International Online Journal of Educational Sciences*, 5(2), 291-301.
- Atar, B. (2007). *Differential item functioning analyses for mixed response data using IRT likelihood-ratio test, logistic regression, and GLLAMM procedures* [Unpublished doctoral dissertation, Florida State University]. http://purl.flvc.org/fsu/fd/FSU_migr_etd-0248
- Baker, F.B. (2001). *The basics of item response theory* (2nd ed.). College Park, (MD): ERIC Clearinghouse on Assessment and Evaluation.
- Baker, F.B., & Kim, S. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). Marcel Dekker.
- Barış-Pekmezci, F. & Şengül-Avşar, A. (2021). A guide for more accurate and precise estimations in simulative unidimensional IRT models. *International Journal of Assessment Tools in Education*, 8(2), 423-453. <https://doi.org/10.21449/ijate.790289>
- Bilir, M.K. (2009). *Mixture item response theory-mimic model: Simultaneous estimation of differential item functioning for manifest groups and latent classes* [Unpublished doctoral dissertation, Florida State University]. <http://diginole.lib.fsu.edu/islandora/object/fsu:182011/datastream/PDF/view>
- Bock, R.D., & Mislevy, R.J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, 6, 431-444. <https://doi.org/10.1177/014662168200600405>
- Bulmer, M.G. (1979). *Principles of statistics*. Dover Publications.
- Bulut, O. & Sünbül, Ö. (2017). R programlama dili ile madde tepki kuramında monte carlo simülasyon çalışmaları [Monte carlo simulation studies in item response theory with the R programming language]. *Journal of Measurement and Evaluation in Education and Psychology*, 8(3), 266-287. <https://doi.org/10.21031/epod.305821>
- Chalmers, R.P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, 48(6), 1-29. <https://doi.org/10.18637/jss.v048.i06>
- Chuah, S.C., Drasgow F., & Luecht, R. (2006). How big is big enough? Sample size requirements for cast item parameter estimation. *Applied Measurement in Education*, 19(3), 241-255. https://doi.org/10.1207/s15324818ame1903_5
- Clarke, E. (2022, December 22). ggbeeswarm: Categorical scatter (violin point) plots. <https://cran.r-project.org/web/packages/ggbeeswarm/index.html>

- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates, Publishers.
- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Wadsworth Group.
- Çelikten, S. & Çakan, M. (2019). Bayesian ve nonBayesian kestirim yöntemlerine dayalı olarak sınıflama indekslerinin TIMSS-2015 matematik testi üzerinde incelenmesi [Investigation of classification indices on Timss-2015 mathematic-subtest through bayesian and nonbayesian estimation methods]. *Necatibey Faculty of Education Electronic Journal of Science and Mathematics Education*, 13(1), 105-124. <https://doi.org/10.17522/balikesirnef.566446>
- De Ayala, R.J. (2009). *The theory and practice of item response theory*. The Guilford Press.
- DeMars, C. (2010). *Item response theory: understanding statistics measurement*. Oxford University Press.
- Demir, E. (2019). *R Diliyle İstatistik Uygulamaları [Statistics Applications with R Language]*. Pegem Akademi.
- Feinberg, R.A., & Rubright, J.D. (2016). Conducting simulation studies in psychometrics. *Educational Measurement: Issues and Practice*, 35, 36-49. <https://doi.org/10.1111/emip.12111>
- Finch, H., & Edwards, J.M. (2016). Rasch model parameter estimation in the presence of a non-normal latent trait using a nonparametric Bayesian approach. *Educational and Psychological Measurement*, 76(4), 662-684. <https://doi.org/10.1177/0013164415608418>
- Fraenkel, J.R., & Wallen, E. (2009). *How to design and evaluate research in education*. McGraw-Hills Companies.
- Gao, F., & Chen, L. (2005). Bayesian or non-Bayesian: A comparison study of item parameter estimation in the three-parameter logistic model. *Applied Measurement in Education*, 18(4), 351-380. https://psycnet.apa.org/doi/10.1207/s15324818ame1804_2
- Goldman, S.H., & Raju, N.S. (1986). Recovery of one- and two-parameter logistic item parameters: An empirical study. *Educational and Psychological Measurement*, 46(1), 11-21. <https://doi.org/10.1177/0013164486461002>
- Hambleton, R.K. (1989). Principles and selected applications of item response theory. In R.L. Linn (Ed.), *Educational Measurement*, (pp.147-200). American Council of Education.
- Hambleton, R.K., & Cook, L.L. (1983). Robustness of item response models and effects of test length and sample size on the precision of ability estimates. In D.J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 31-49). Vancouver.
- Hambleton, R.K., & Jones, R.W. (1993). An NCME instructional module on comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, 12(3), 38-47. <https://doi.org/10.1111/j.1745-3992.1993.tb00543.x>
- Hambleton, R.K., & Swaminathan, H. (1985). *Item response theory: Principals and applications*. Kluwer Academic Publishers.
- Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of item response theory*. Sage Publications Inc.
- Harwell, M., & Janosky, J. (1991). An empirical study of the effects of small datasets and varying prior variances on item parameter estimation in BILOG. *Applied Psychological Measurement*, 15, 279-291. <https://doi.org/10.1177/014662169101500308>
- 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. <https://doi.org/10.1177/014662169602000201>
- Hoaglin, D.C., & Andrews, D.F. (1975). The reporting of computation-based results in statistics. *The American Statistician*, 29, 122-126. <https://doi.org/10.2307/2683438>

- 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, 249-260. <https://psycnet.apa.org/doi/10.1177/014662168200600301>
- Karadavut, T. (2019). The uniform prior for bayesian estimation of ability in item response theory models. *International Journal of Assessment Tools in Education*, 6(4), 568-579. <https://dx.doi.org/10.21449/ijate.581314>
- Kıbrıslıoğlu Uysal, N. (2020). *Parametrik ve Parametrik Olmayan Madde Tepki Modellerinin Kestirim Hatalarının Karşılaştırılması [Comparison of estimation errors in parametric and nonparametric item response theory models]* [Unpublished doctoral dissertation, Hacettepe University]. <http://hdl.handle.net/11655/22495>
- Kirisci, L., Hsu, T.C., & Yu, L. (2001). Robustness of item parameter estimation programs to assumptions of unidimensionality and normality. *Applied Psychological Measurement*, 25(2), 146-162. <https://doi.org/10.1177/01466210122031975>
- Kolen, M.J. (1985). Standard errors of Tucker equating. *Applied Psychological Measurement*, 9(2), 209-223. <https://doi.org/10.1177/014662168500900209>
- Kothari, C.R. (2004). *Research methodology: methods and techniques* (2nd ed.). New Age International Publishers.
- Köse, İ.A. (2010). *Madde Tepki Kuramına Dayalı Tek Boyutlu ve Çok Boyutlu Modellerin Test Uzunluğu ve Örneklem Büyüklüğü Açısından Karşılaştırılması [Comparison of Unidimensional and Multidimensional Models Based On Item Response Theory In Terms of Test Length and Sample Size]* [Unpublished doctoral dissertation]. Ankara University, Institute of Educational Sciences.
- Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *frontiers in Psychology*, 4(863), 1-12. <https://doi.org/10.3389/fpsyg.2013.00863>
- Lenth, R.V. (2022, December). emmeans: Estimated marginal means, aka Least-Squares Means. <https://cran.r-project.org/web/packages/emmeans/index.html>
- Lim, H., & Wells, C.S. (2020). irtplay: An R package for online item calibration, scoring, evaluation of model fit, and useful functions for unidimensional IRT. *Applied psychological measurement*, 44(7-8), 563-565. <https://doi.org/10.1177/0146621620921247>
- Linacre, J.M. (2008). *A user's guide to winsteps ministep: rasch-model computer programs*. <https://www.winsteps.com/winman/copyright.htm>
- Lord, F.M. (1968). An analysis of the verbal scholastic aptitude test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement*, 28, 989-1020. <https://doi.org/10.1177/001316446802800401>
- Lord, F.M. (1980). *Applications of item response theory to practical testing problems*. Lawrence Erlbaum Associates.
- Lord, F.M. (1983). Unbiased estimators of ability parameters, of their variance, and of their parallel forms reliability. *Psychometrika*, 48, 233-245. <https://doi.org/10.1007/BF02294018>
- Martin, A.D., & Quinn, K.M. (2006). Applied Bayesian inference in R using MCMCpack. *The Newsletter of the R Project*, 6(1), 2-7.
- Martinez, J. (2017, December 1). bairt: Bayesian analysis of item response theory models. <https://cran.nexr.com/web/packages/bairt/index.html>
- Maydeu-Ovives, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika*, 71, 713-732. <https://doi.org/10.1007/s11336-005-1295-9>
- Meyer, D. (2022, December 1). e1071: Misc functions of the department of statistics, Probability Theory Group (Formerly: E1071), TU Wien. <https://CRAN.R-project.org/package=e1071>

- Mislevy, R.J. (1986). Bayes modal estimation in item response models. *Psychometrika*, 51, 177-195. <https://doi.org/10.1007/BF02293979>
- MoNE (2022). *Sınavla Öğrenci Alacak Ortaöğretim Kurumlarına İlişkin Merkezî Sınav Başvuru ve Uygulama Kılavuzu [Central Examination Application and Administration Guide for Secondary Education Schools to Admit Students by Examination]*. Ankara: MoNE [MEB]. <https://www.meb.gov.tr/2022-lgs-kapsamindaki-merkez-sinav-kilavuzu-yayimlandi/haber/25705/tr>
- Morris, T.P., White, I.R., & Crowther, M.J. (2017). Using simulation studies to evaluate statistical methods. *Tutorial in Biostatistics*, 38(11), 2074-2102. <https://doi.org/10.1002/sim.8086>
- Orlando, M. (2004, June). Critical issues to address when applying item response theory models. *Paper presented at the conference on improving health outcomes assessment*, National Cancer institute, Bethesda, MD, USA.
- Pekmezci Barış, F. (2018). *İki Faktör Modelde (Bifactor) Diklik Varsayımının Farklı Koşullar Altında Sinanması [Investigation Of Orthogonality Assumption In Bifactor Model Under Different Conditions]* [Unpublished doctoral dissertation]. Ankara University, Institute of Educational Sciences, Ankara.
- Ree, M.J., & Jensen, H.E. (1980). Effects of sample size on linear equating of item characteristic curve parameters. In D.J. Weiss (Ed.), *Proceedings of the 1979 computerized adaptive testing conference*. (pp. 218-228). Minneapolis: University of Minnesota. <https://doi.org/10.1016/B978-0-12-742780-5.50017-2>
- Reise, S.P., & Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. *Journal of Educational Measurement*, 27, 133-144. <https://doi.org/10.1111/j.1745-3984.1990.tb00738.x>
- Revelle, W. (2022, October). psych: Procedures for psychological, psychometric, and personality research. <https://cran.r-project.org/web/packages/psych/index.html>
- Robitzsch, A. (2022). sirt: Supplementary item response theory models. <https://cran.r-project.org/web/packages/sirt/index.html>
- Samejima, F. (1993a). An approximation for the bias function of the maximum likelihood estimate of a latent variable for the general case where the item responses are discrete. *Psychometrika*, 58, 119-138. <https://doi.org/10.1007/BF02294476>
- Samejima, F. (1993b). The bias function of the maximum likelihood estimate of ability for the dichotomous response level. *Psychometrika*, 58, 195-209. <https://doi.org/10.1007/BF02294573>
- Sarkar, D. (2022, October). lattice: Trellis graphics for R. R package version 0.20-45, URL <http://CRAN.R-project.org/package=lattice>.
- SAS Institute (2020). Introduction to Bayesian analysis procedures. In *User's Guide Introduction to Bayesian Analysis Procedures*. (pp. 127-161). SAS Institute Inc., Cary, (NC), USA.
- Sass, D., Schmitt, T., & Walker, C. (2008). Estimating non-normal latent trait distributions within item response theory using true and estimated item parameters. *Applied Measurement in Education*, 21(1), 65-88. <https://doi.org/10.1080/08957340701796415>
- Seong, T.J. (1990). Sensitivity of marginal maximum likelihood estimation of item and ability parameters to the characteristics of the prior ability distributions. *Applied Psychological Measurement*, 14(3), 299-311. <https://psycnet.apa.org/doi/10.1177/014662169001400307>
- Singmann, H. (2022, December). afex: Analysis of factorial experiments. <https://cran.r-project.org/web/packages/afex/afex.pdf>
- Soysal, S. (2017). *Toplam Test Puanı ve Alt Test Puanlarının Kestiriminin Hiyerarşik Madde Tepki Kuramı Modelleri ile Karşılaştırılması [Comparison of Estimation of Total Score and Subscores with Hierarchical Item Response Theory Models]* [Unpublished doctoral dissertation]. Hacettepe University, Institute of Educational Sciences, Ankara.

- Stone, C.A. (1992). Recovery of marginal maximum likelihood estimates in the two parameter logistic response model: An evaluation of MULTILOG. *Applied Psychological Measurement, 16*(1), 1-16. <https://doi.org/10.1177/014662169201600101>
- Swaminathan, H., & Gifford, J.A. (1986). Bayesian estimation in the three-parameter logistic model. *Psychometrika, 51*, 589-601. <https://doi.org/10.1007/BF02295598>
- Şahin, A., & Anıl, D. (2017). The effects of test length and sample size on item parameters in item response theory. *Educational Sciences: Theory & Practice, 17*, 321-335. <http://dx.doi.org/10.12738/estp.2017.1.0270>
- Tabachnick, B.G., & Fidell, L.S. (2014). *Using multivariate statistics* (6th ed.). Pearson New International Edition.
- Thissen, D., & Wainer, H. (1983). Some standard errors in item response theory. *Psychometrika, 47*, 397-412. <https://doi.org/10.1007/BF02293705>
- Thorndike, L.R. (1982). *Applied Psychometrics*. Houghton Mifflin Co.
- Van de Schoot, R., & Depaoli, S. (2014). Bayesian analyses: Where to start and what to report. *The European Health Psychologist, 16*(2), 75-84.
- Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag. <https://doi.org/10.1007/978-0-387-98141-3>
- Wright, B.D., & Stone, M.H. (1979). *Best test design*. Mesa Press
- Yen, W.M. (1984). Effects of local item dependence on the fit and equating performance of the three parameter logistic model. *Applied Psychological Measurement, 8*, 125-145. <https://doi.org/10.1177/014662168400800201>

APPENDIX**#ÖNSEL (PRIOR) SCRIPT BLOCK***

```
# Generation of necessary prior distributions and data sets according to simulation conditions
```

```
library(psych)
```

```
library(e1071)
```

```
library(mirt)
```

```
#I: number of items
```

```
#N: number of individuals
```

```
#M: number of parameters
```

```
#D: distribution state
```

```
prior <- function (I, N, M=c("2PL", "3PL"), D=c("normal", "left-skewed", "right-  
skewed", "leptokurtic", "platykurtic"))
```

```
{
```

```
a <- rlnorm(I, meanlog = 0.3, sdlog = 0.2)
```

```
b <- rnorm(I, mean = 0, sd = 1)
```

```
c <- runif(I, min = 0.01, max = 0.25)
```

```
if (D=="normal") {k <- as.matrix(rnorm(N, mean = 0, sd = 1))}
```

```
else if (D==" left-skewed") {k <- as.matrix(c(rnorm(N*86/100, 2, 1)), runif(N*7/100, min = -  
5, max = -4), runif(N*7/100, min = -4, max = -3))}
```

```
else if (D=="right-skewed ") {k <- as.matrix(c(rnorm(N*86/100, -2, 1)), runif(N*7/100, min  
= 3, max = 4), runif(N*7/100, min = 4, max = 5))}
```

```
else if (D=="leptokurtic ") {k <- as.matrix(c(rnorm(N*3/100, -1, 100), rnorm(N*94/100, 0,  
0.00001), rnorm(N*3/100, 1, 100)) )}
```

```
else if (D=="platykurtic") {k <- as.matrix(c(rnorm(N*40/100, 0, 1)), runif(N*30/100, min = -  
3, max = -1), runif(N*30/100, min = 1, max = 3))}
```

```
if (M=="2PL")
```

```
{dat <- as.data.frame(simdata(a = a, d = b, N = N, itemtype = "dich", Theta = k))
```

```
model2pl <- mirt(data = dat, 1, itemtype = "2PL", SE = TRUE, verbose = FALSE, technical =  
list(NCYCLES = 10000))
```

```
irt.parameters <- as.data.frame(coef(model2pl, simplify = TRUE)$items)
```

```
bias.a <- mean(irt.parameters[,1]-a)
```

```
bias.b <- mean(irt.parameters[,2]-b)
```

```
rmse.a <- sqrt(mean((irt.parameters[,1]-a)^2))
```

```
rmse.b <- sqrt(mean((irt.parameters[,2]-b)^2))
```

```
fit2pl <- M2(model2pl)
```

```
M2 <- fit2pl$M2
```

```
p <- fit2pl$p
```

```
data <- list(dat, bias.a, rmse.a, bias.b, rmse.b, M2, p, k)
```

```
print(data)}

else if (M=="3PL")

{
dat <- as.data.frame(simdata(a = a, d = b, guess = c, N = N, itemtype = "dich", Theta = k))

model3pl <- mirt(data = dat, 1, itemtype = "3PL", SE = TRUE, verbose = FALSE, technical =
list(NCYCLES = 10000))

parameters <- as.data.frame(coef(model3pl, simplify = TRUE)$item)
bias.a <- mean(parameters[,1]-a)
bias.b <- mean(parameters[,2]-b)
bias.c <- mean(parameters[,3]-c)
rmse.a <- sqrt(mean((parameters[,1]-a)^2))
rmse.b <- sqrt(mean((parameters[,2]-b)^2))
rmse.c <- sqrt(mean((parameters[,3]-c)^2))

fit3pl <- M2(model3pl)
M2 <- fit3pl$M2
p <- fit3pl$p

data <- list(dat, bias.a, rmse.a, bias.b, rmse.b, bias.c, rmse.c, M2, p, k)
print(data)} }
```

**The codes of Bulut and Sünbiül (2017) were used in some parts of this function.*