REID (Research and Evaluation in Education), 11(1), 2025, 101-111
Available online at: http://journal.uny.ac.id/index.php/reid

A comparison of the stability of ability parameter estimation based on the maximum likelihood and Bayesian estimation: A case study of dichotomous scoring test results

Faradila Ilena Putri*[1]; Heri Retnawati[1]; Elena Kardanova[2]
[1]Universitas Negeri Yogyakarta, Indonesia
[2]National Research University, Higher School of Economics (HSE University), Russian Federation
*Corresponding Author. E-mail: dilafara152@gmail.com

ARTICLE INFO

Article History
Submitted:
August 28, 2025
Revised:
September 19, 2025
Accepted:
September 24, 2025

Keywords
ability estimation; Bayes method; maximum likelihood method; item response theory; dichotomous scoring test

Scan Me:

ABSTRACT

This research is related to Item Response Theory (IRT), which is essential for determining the best method for estimating participants' abilities on a test measuring English listening ability. This study aims to (1) determine the characteristics of the test device measuring English listening ability, (2) determine the effect of the length of the test on the stability of the ability estimation using the maximum likelihood (ML) method, (3) determine the effect of test length on the stability of the ability estimation using the Bayes method, and (4) compare the stability of the ability estimate between ML and Bayes. This research is an exploratory descriptive study using a simulation approach. The best model is selected to generate data. The result of the generation is the actual ability (θ) and the participant's response, which is estimated with the maximum likelihood and Bayes, which produces the estimated ability with 10 replications, and is compared with calculating the MSE (mean square error). The method with a smaller MSE is stable and has a better estimation method. The results show that (1) the 2PL model is the best, (2) the length of the test affects the stability of the ability estimation in the ML method and the most stable case when the test contains 46 items, (3) the length of the test affects the stability of the ability estimate in the Bayes method and it is most stable when the test contains 46 items, and (4) the Bayes method is better and more accurate for estimating ability.

To cite this article (in APA style):
Putri, F. I., Retnawati, H., & Kardanova, E. (2025). A comparison of the stability of ability parameter estimation based on the maximum likelihood and Bayesian estimation: A case study of dichotomous scoring test results. *REID (Research and Evaluation in Education)*, *11*(1), 101-111. https://doi.org/10.21831/reid.v11i1.89463

INTRODUCTION

Tests in education for language development, specifically to measure language abilities, are conducted to assess skills related to specific languages. In the language ability test, a test set is made in the form of items to measure language skills such as competency and intelligence, where competency and intelligence are latent abilities that cannot be measured directly, so there is a need for visible indicators to form instruments that are then used to collect test-taker responses. In education, participant response scoring employs a dichotomous format, where responses are categorized as true or false, across the test set, specifically, the test items. It is necessary to estimate the item parameters to measure how well the test device measures the language abilities of the test-takers. Therefore, their ability can be assessed with the test-taker's response from the existing test kit. The ability parameter is a measure or criterion of the ability of someone or something to achieve a goal or perform a particular task (Retnawati, 2014). Estimation of ability parameters can be used to measure the characteristics of test-takers by reading the values.

ISSN: 2460-6995 (Online)

Two methods are often used to estimate ability parameters, namely, the maximum likelihood (ML) and the Bayesian methods (Retnawati, 2014). The ML method is a method for determining the maximum likelihood function, while the Bayes method is based on the average of the posterior distribution (Bock & Aitkin, 1981). The ML method is widely used in IRT because it is asymptotically unbiased and efficient with large sample sizes, and it provides straightforward and precise estimates. In contrast, the Bayes method is advantageous when incorporating prior knowledge, handling small or noisy data, or modeling complex structures, as it provides flexible estimation with uncertainty measures. While ML is simpler and efficient for large datasets, Bayesian estimation is more robust in uncertain conditions. However, it remains unclear which method is more effective in practice, as effectiveness in IRT depends on how accurately each method estimates ability. The effectiveness of a method can be evaluated based on several factors, including accuracy of measurement, reliability, validity, computational efficiency, applicability in different contexts, and users and measurement practitioners who use Item Response Theory often argue about which of the two methods is more effective (Retnawati, 2015), because the two methods have their advantages and disadvantages. The weaknesses and strengths of the two methods make debate about which method is more effective so there is still a need for a lot of research or studies regarding the comparison of the two methods, which was previously carried out by Retnawati (2015) regarding comparing the two methods namely ML and Bayes related to Item Response Theory, where the Bayes method used is Expected A posteriori (EAP). The EAP method is an estimated a posteriori that refers to the expected value of the posterior probability distribution of latent trait values for certain cases (de Ayala, 2010) and is carried out by modifying the likelihood function. The stability of the two methods will be reviewed based on the test length (20, 25, 30, 35, 40, and 46 items), where the test length refers to the number of items or test sets. The lengths of the tests used are 20, 25, 30, 35, and 40 minutes, as these durations are commonly used in educational scoring. The study indeed focuses on a test device for measuring English listening ability. Although the test length is known, our consideration of different cases aims to explore potential variations in measurement precision. Additionally, in our simulation process, we acknowledge the importance of incorporating other parameters, such as item characteristics and ability distributions, to ensure realistic and meaningful data analysis.

This research was conducted using IRT on dichotomous data with a Monte Carlo simulation, a method that generates random numbers to obtain numerical solutions for complex problems (Harwell et al., 1996). Monte Carlo is useful for predicting errors from empirical distributions (Hammersley & Handscomb, 1964), and in this study, 10 data replications were generated, as recommended by Harwell et al. (1996). The simulation used item parameters of difficulty (a) and discrimination (b) from the original English listening ability test data to produce both actual and estimated ability values, whose accuracy was evaluated using Mean Square Error (MSE). The method with the smaller MSE was considered more effective (Retnawati, 2015), with test length also plotted against estimation accuracy. Retnawati (2015), for example, compared ML and Bayes methods using National Examination data with varying test lengths (15–30 items) and sample sizes (500–1,500). Her findings showed unstable MSE values, but with 1,500 participants, both methods produced similar accuracy, while ML was more accurate for longer tests (25–30 items). Similarly, Yendra and Noviadi (2015) found that ML outperformed Bayes in parameter estimation of exponential distributions, as shown by lower AIC values.

Based on the aforementioned background, simulation research is needed on the stability (accuracy) of ability estimation between ML and Bayesian methods. This comparison is important because researchers require precise ability-parameter estimation to decide which method is most effective. Accordingly, this study addresses the following research questions: (1) Which IRT model (Rasch, 1PL, 2PL, or 3PL) best fits the English listening ability test data? (2) Do the assumptions of unidimensionality, parameter invariance, and local independence hold for the selected IRT model? (3) How does test length (20, 25, 30, 35, 40, and 46 items) affect the

accuracy of ability estimation for ML and Bayes methods? and (4) Which estimation method (ML or Bayes) demonstrates greater stability (lower MSE) across test lengths? The answer from this research can be considered by experts or researchers regarding the use of parameter estimation methods to be used.

METHOD

Data Description

This study employs a simulation approach, specifically Monte Carlo simulation, to compare the stability of ability estimation between the maximum likelihood (ML) and Bayesian methods using dichotomous data from Item Response Theory. The data used in this study are secondary data from test kits measuring Pro-TEFL English listening ability in 2021 at a university in Yogyakarta. The data is the dichotomous scoring data from the Pro-TEFL English listening ability test kit. The population data in this study consists of 3,042 test-taker responses who answered one of the item identities (IDs) in the English listening proficiency Pro-TEFL test.

Data Analysis Procedure

The data analysis followed the following stages: (1) preparing test device data for measuring English listening ability, (2) testing the suitability of the fittest model with the most items using the Rasch model, 1PL model, 2PL model, and 3PL model using the Chi-Square test statistic, (3) testing the assumptions of the Item Response Theory, namely the assumption of unidimensionality, parameter invariance, and local independence with the help of R using the PCA, get_eigenvalue and fviz_eig functions in the factoextra package (Kassambara & Mundt, 2016) and factorMineR (Lê et al., 2008), (4) estimation of item parameters (test device characteristics) which produce discrimination power parameters (a) and level of difficulty (b) and ability parameter estimates which produce ability values from the original data of English listening ability test kits based on the best model, (5) Monte Carlo simulation to generate data, used to predict the error obtained from the empirical distribution function of the samples obtained (Hammersley & Handscomb, 1964). Monte Carlo simulation is used in Item Response Theory to provide information on how valid this method can be applied to a data set, (6) estimation of ability parameters using the Maximum Likelihood (ML) and Bayes methods, namely EAP on the original data of English listening ability test kits based on the best model, and (7) analysis of the results of the Monte Carlo study, in accordance with the objectives of the study, will compare the results of the MSE between the two methods, namely ML and Bayes.

FINDINGS AND DISCUSSION

Findings

Data processing was done using RStudio software. The data from language proficiency test kits were estimated using model fit tests with the Rasch, 1PL, 2PL, and 3PL models, as well as item characteristic curve graphs. The results obtained were compared based on the number of items that matched. Then, after obtaining the best model, the researchers tested the assumptions of unidimensionality, parameter invariance, and local independence. The assumptions that were met were used to determine the characteristics of the test set and the characteristics of the test-takers. After the assumptions on the best model were met, the best data model was used to generate data with Monte Carlo simulations using the estimated item parameter model and the distribution of participants' abilities with 10 replications using the R program according to the length of the test or item that has been determined, namely 20, 25, 30, 35, 40, or 46 items. The generated data represents both the participant's actual ability and their response. Then, the participants' responses were analyzed with maximum likelihood and Bayes for each replication

according to the number of test items to obtain the ability value estimation results. The results of the two methods were compared by calculating the MSE. At each test length used, the MSE was calculated as the average of 10 replications. The methods with a smaller MSE were said to be stable or had stability in their estimating ability, and were considered better estimation methods.

Model Fitness

The selection of the best model can be determined by testing the model's fit using Yen's Q1 method and examining the Item Characteristics Curve graph. The value of Q_1 will be compared with the X2 Table with degrees of freedom of $10 \times (J - 1) -$. Thus, the Rasch model on dichotomous data has a degree of freedom of 9. An item can be declared unsuitable if $Q_1 > X^2_{0,05(9)} = 16,92$ for the Rasch Model because in this model, it only estimates parameter b, $Q_1 > X^2_{0,05(9)} = 15,51$ for 1PL and 2PL models, and $Q_1 > X^2_{0,05(9)} = 14,07$ for the 3PL model because this model estimates three parameters *(a, b*, g), as presented in Table 1.

Table 1. Summary of Item Fits to Model

Category	Rasch Model	1PL Model	2PL Model	3PL Model
Fit	11	8	15	13
Not Fit	35	38	31	33

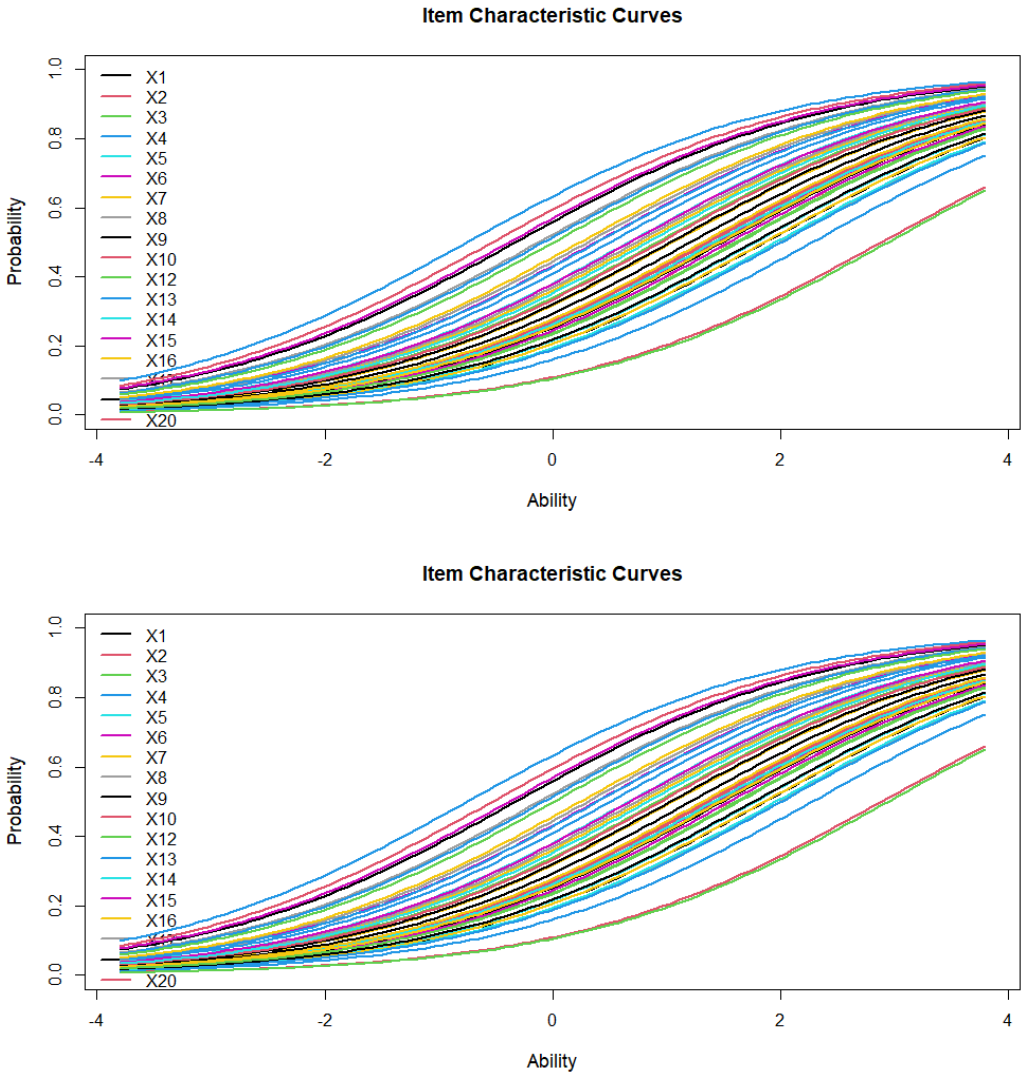

Figure 1. Plot ICC of Rasch and 1PL Models

The Item Characteristics Curve graph can be used to select the best model. Figure 1 and Figure 2 present an ICC plot for the Rasch, 1PL, 2PL, and 3PL models.

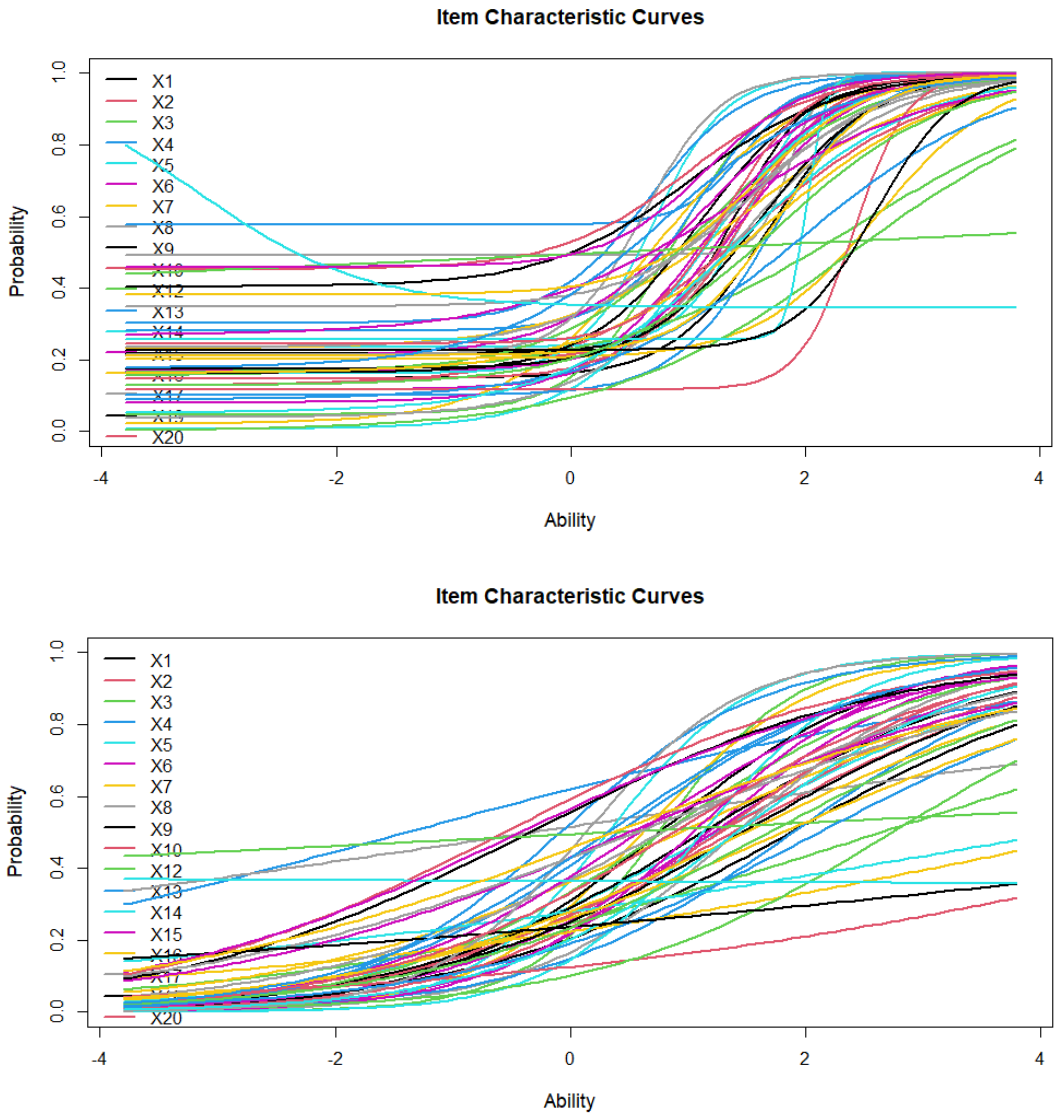

Figure 2. Plot ICC of 2PL and 3PL Models

Table 1 shows the summary of the model fit test. Thus, from the model fit test, the 2PL model is the best model for the English listening ability test data, since it has the most items that match the data. However, many do not fit the 2PL model; far more fit the 2PL model than the 3PL model. In Figure 2, the plot of the Item Characteristics Curve of the 2PL model, although some items produce graphs that do not follow the normal ogif and some do not form an S-curve, most items have a graphical shape that follows the normal ogif, that is, when the graph is in the form of an S, with these results the ICC plot of the 2PL model is still acceptable.

Figure 2 shows that the ICC 3PL plot produces a graph in which most items do not follow the normal ogif and do not form an S curve. Therefore, using the model fit test and graphs, the researchers determined the 2PL model as the best model because it has the largest number of suitable items, and the resulting ICC plot is also quite good. In this study, the 2PL model was identified as the best model for research comparing the stability of ability estimation between the ML and Bayes methods, considering the test length on the data from the test kit measuring English listening ability.

Item Response Theory Assumption

Unidimensionality

Figure 3 shows that the presentation of the variance from the first to the second dimension has decreased steeply by 11.2% and the first dimension is very dominant compared to the other dimensions and there are elbow points from the first to the second dimension, so it can be concluded that the test device measures only one dimension and it is called unidimensional, and the assumption of unidimensionality is met.

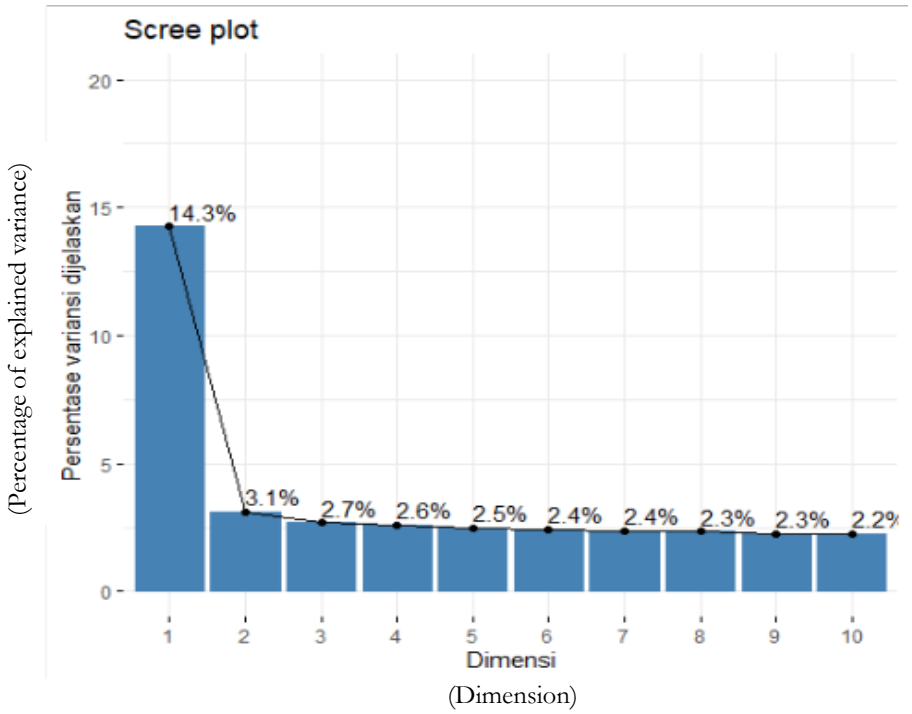

Figure 3. Scree Plot on Data

Parameter Invariance

The x represents ability levels (θ) or item parameters such as difficulty and discrimination. At the same time, the y may indicate the probability of a correct response, parameter estimates, or differences in estimates across conditions. If the graph presents item characteristic curves, it illustrates the probability of a proper response as a function of ability. If it compares parameter estimates, it shows the stability of item parameters across different samples or estimation methods.

In IRT, parameter invariance means that item parameters should remain stable regardless of the examinee group used for estimation. If significant variations in parameter estimates occur across different groups, this suggests a violation of invariance, indicating possible model misfit. Understanding these aspects helps assess the validity of the model and the reliability of estimated parameters.

Figure 4 shows that the points are spread around the line and follow a straight line. Thus, it can be concluded that the assumptions of invariance of the discrimination power item parameters (a) and level of difficulty (b) in the English listening ability test kit data are fulfilled.

Figure 5 shows that the points also spread around the line and follow a straight line. Thus, it can be concluded that the assumption of the ability invariance parameter on the test device data to measure language ability is fulfilled.

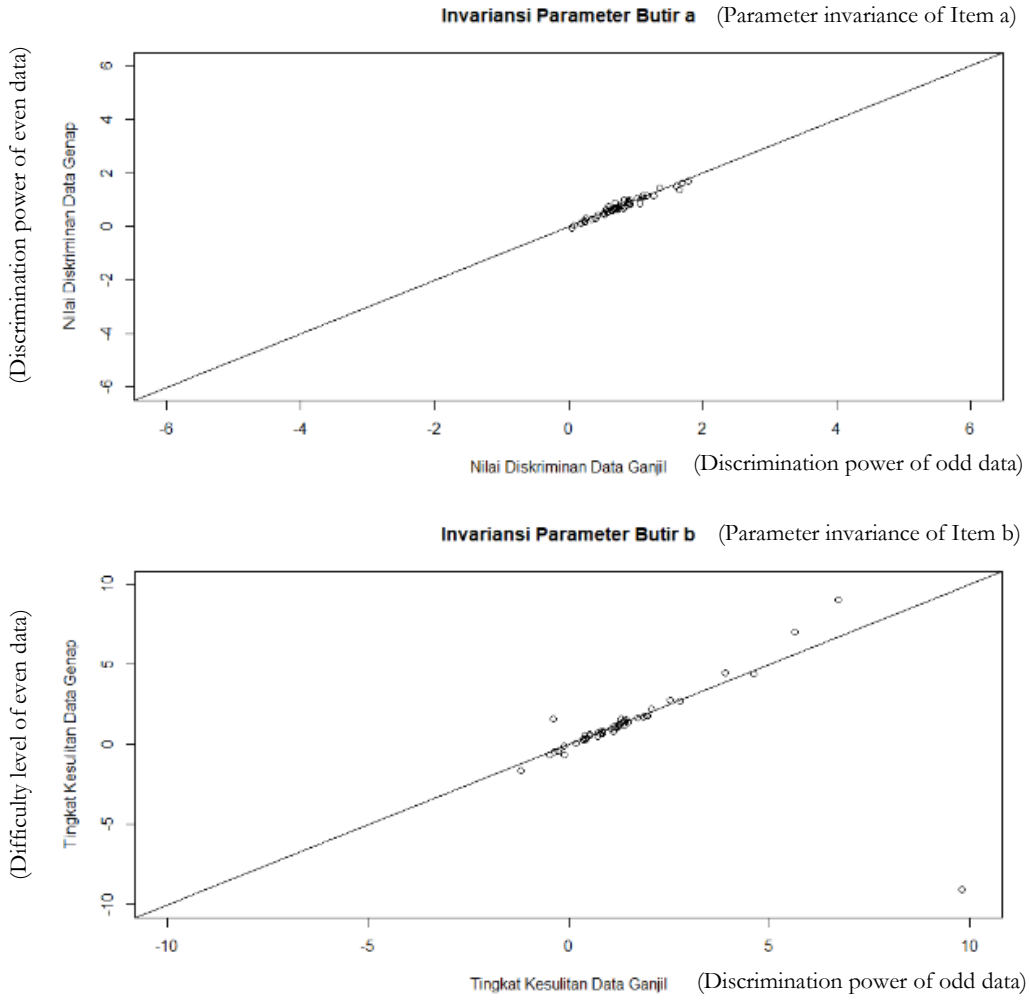

Figure 4. Parameter Invariance of Item Discrimination Power (a) and Difficulty Level (b)

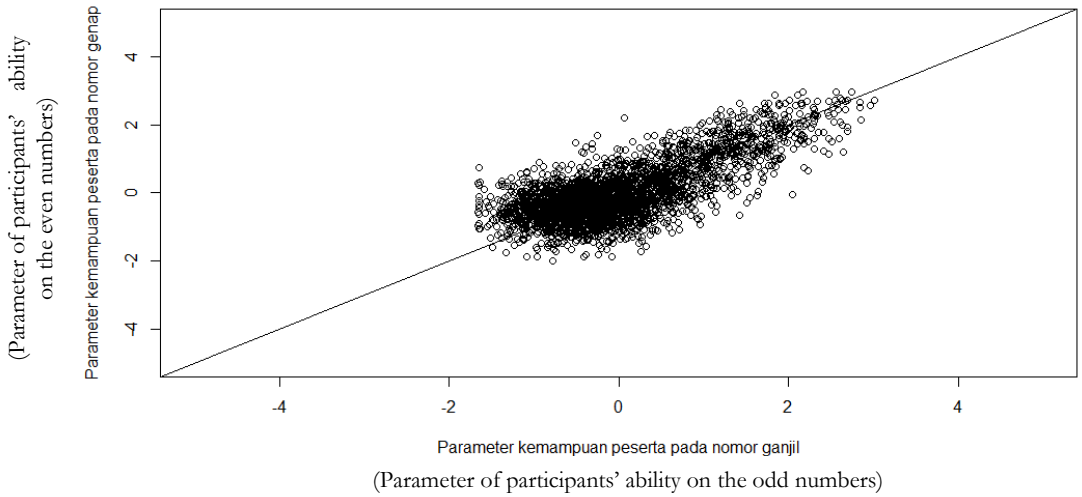

Figure 5. Parameter Invariance of Ability

Local Independency

Since the unidimensional assumption has been met, the local independence assumption has also been satisfied (Hambleton et al., 1991). The local independence test is fulfilled if the participants' answers to one item do not affect their answers to other items.

Characteristics of the Language Competence Test Battery

The quality of a test is typically evaluated based on several key psychometric properties, elaborated as follows. "Valid" refers to measuring the intended language skills accurately (content, construct, and criterion validity). "Reliable" refers to producing consistent results across different conditions (internal consistency, test-retest, and inter-rater reliability). "Well-balanced in difficulty" includes easy to difficult items to differentiate proficiency levels. "Discriminative" effectively distinguishes between test-takers of varying abilities. "Varied in format" can be objective (multiple-choice), subjective (essays, interviews), or adaptive (CAT). "Comprehensive" assesses listening, reading, writing, and speaking skills. "Practical" refers to being easy to administer, score, and interpret, with reasonable cost and time requirements.

Monte Carlo Simulation

The Monte Carlo simulation process began by estimating the items and abilities by using the most suitable model on the data of the English listening ability test kit, namely the 2PL model. The item estimation of the 2PL model produced parameters for difficulty level (b) and discriminating power (a), as well as an estimate of the model's ability 2PL, and obtained a capability value. From the estimation results, the model generated the figure and table data. Furthermore, with the result parameters of the best real data model, Item parameters (such as difficulty, discrimination, and guessing) define how test items function within an IRT model that is simulate realistic test conditions before applying the model to actual data, evaluate model performance under controlled conditions, and compare different estimation methods to identify which provides the most accurate results, the data generated item parameters on standard IRT models: difficulty level (b), discriminating power (a), and guessing parameters (c) for 3PL model. The relationship simulation and real data in this study serve as a controlled experiment to test the performance of estimation methods under ideal conditions. Unlike real test data, which may contain noise or missing responses, simulation allows for a deeper understanding of the theoretical properties of the estimation techniques used. If the simulation results align with real data analysis, this indicates that the applied estimation method is highly reliable and can be effectively used in real testing scenarios.

The data were generated using a normal and uniform probability distribution with the help of the R-studio program using the *rnorm* and *runif* functions. Item parameters, namely the values of *a* and *b* and the distribution of abilities according to each length of the test resulting from the original data, namely a test kit measuring English listening ability, where the values of *a* and *b* were generated using a uniform probability distribution, the minimum, maximum, mean and SD values were used on the ability distribution using the normal probability distribution. This study uses test length as a variable (n=20, n=25, n=30, n=35, n=40, and n=46). The results of the data generation from the length of the test used produce the participants' abilities, which are considered to be the actual abilities, and the participants' responses, which are then estimated by ML and Bayes, as presented in Table 2 and Figure 6.

Table 2. MSE of Item Parameter Estimation by ML and Bayes across Test Lengths

	MSE	
Items	**ML**	**Bayes**
20	9,894.76	338.80
25	5,581.61	331.47
30	4,772.32	330.00
35	4,322.31	324.04
40	1,406.46	323.49
46	673.54	306.99

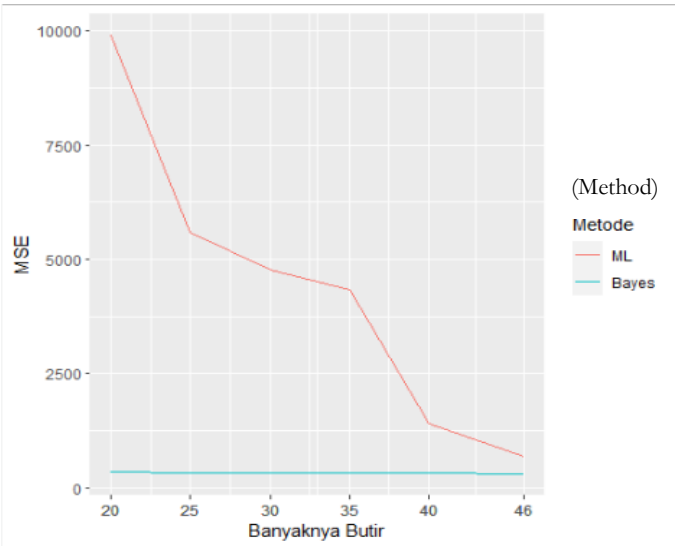

Figure 6. MSE of Item Parameter Estimation by ML and Bayesian Methods Across Test Lengths

The Stability of Ability Estimation with the ML Method by Looking at the Effect of Test Length (20, 25, 30, 35, 40, and 46 Items)

The figure for calculating the MSE ability using the ML method shows that the graph results are decreasing because the length of the 20-item to 46-item test decreases. Table 2 shows that the number of test lengths affects the estimation of ability parameters because the resulting MSE value gets smaller with the increasing number of test lengths used, and the longer the test, the more accurate the estimation of ability parameters will be. The ML method shows that the results, as indicated by the graph presented in Figure 6, tend to decrease.

The Stability of Ability Estimation with the ML Method by Looking at the Effect of Test Length (20, 25, 30, 35, 40, and 46 Items)

The figure for calculating the MSE ability with the Bayes method shows that the calculation of MSE ability with the Bayes method is quite stable, as indicated by the downward-sloping graph. The table shows that the longer the test, the smaller the MSE value, indicating that the test length affects the stability of ability estimation in the Bayes method. Furthermore, the longer the test, the more accurate the ability parameter estimates become.

Comparison of the Results of the Ability Estimation Stability with ML and Bayes Methods

The results obtained show that the stability results of the ability estimation using the Bayes method have smaller MSE results compared to the ability estimation using the ML method, and the Bayes method graph shows a more stable graph because the MSE results on the Bayes method are in the range of 300 so that the graph tends to be sloping while ML tends to decrease steeply because the MSE results obtained range from 600-9,000. The stability of MSE in both methods is Similar in the study "Comparison of Latent Ability Estimation between Maximum Likelihood and Bayesian Methods".

Discussion

This study aims to compare the stability (accuracy) of ability estimation between the maximum likelihood and Bayesian methods by reviewing the test length variable. The method with the lowest stability (accuracy) results is the best method for estimating abilities.

Based on the results obtained in this study using data from language ability test kits with listening question types, the length of the test affects the stability of ability estimation. This is in

line with the research of Falani and Kumala (2017), which suggests that the length of the test affects the stability of ability parameter estimation. The estimation results obtained show that the longer the test used, the more accurate the estimation of ability parameters will be. The results obtained show that for the maximum likelihood method, a test containing 30 items produces fairly good accuracy, and for the Bayes method, tests containing 25 and 20 items are also quite good, so that at the test length, the two methods have the same good estimation accuracy results. Mahmud et al. (2016) also found that simply increasing the number of test items does not necessarily lead to lower variance when using MLE. In contrast, EAP consistently showed more stable variance across different test lengths, which reinforces the findings of this study. These results are in line with Retnawati (2015) that a test containing 25 and 30 items with 1,500 test-takers obtained the same good estimation accuracy results.

This comparison of the stability of the ability estimation between the ML and Bayes methods yields the result that the Bayes method is a better estimation method to use because it has a smaller MSE, and this is in accordance with a study by Hikamudin (2017) that the Bayes method produces smaller and more accurate MSE, compared to the MSE of the ML method, and the study conducted by Insuk (2007), who used empirical data and simulation data resulting in that estimation with the Bayes method was better than estimation with ML method in all conditions related to item parameter estimation. This finding is also in line with the theoretical insights in IRT, which suggest that EAP tends to yield more stable ability estimates, especially when dealing with small samples or extreme response patterns. In contrast, MLE is generally more effective for large samples but can become less reliable when all responses are either correct or incorrect (Mahmud, 2017). This finding differs from the results of the study by Yendra and Noviadi (2015) that by testing the adequacy of the AIC of the parameters used, it was concluded that the ML method was better than the Bayes method in estimating.

CONCLUSION

Based on the results of research related to the stability of item parameter estimates and ability in the dichotomy data of a test kit measuring English listening ability in 2021, it can be concluded that the most suitable model for estimating parameters in the listening test kit data is the 2PL model. The length of the test is a variable that can affect the stability of ability parameter estimation in the ML method and the Bayes method. The most stable case is when the test contains 46 items, compared to tests containing 20, 25, 30, 35, or 40 items. This method is a more accurate and reliable method for estimation because it has the smallest MSE value and a more stable graph.

Suggestions for further research include comparing the stability of ability parameter estimation in the ML and Bayes methods using other question identity data in listening question types. Further research is recommended to use the stability of ability estimates other than MSE, such as RMSEA. It is also recommended that further research utilise other variables that can affect the stability of the estimate, such as the number of participants, and employ alternative methods, including non-Item Response Theory approaches, such as linear regression.

DISCLOSURE STATEMENT

The authors declare that they have no conflicts of interest to disclose.

FUNDING STATEMENT

The research does not receive funding.

ETHICS APPROVAL

This study complied with ethical standards and data privacy regulations. All data were used solely for research, kept confidential, and handled securely.

REFERENCES

Kassambara, A., & Mundt, F. (2016). *Package 'factoextra.'* https://doi.org/10.32614/CRAN.package.factoextra

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters. *Psychometrika*, *46*(4), 443–459. https://doi.org/10.1007/BF02293801

de Ayala, R. J. (2010). *The theory and practice of item response theory*. Guilford Press.

Falani, I., & Kumala, S. A. (2017). Kestabilan estimasi parameter kemampuan pada model logistik item response theory ditinjau dari panjang tes. *SAP (Susunan Artikel Pendidikan)*, *2*(2), . https://doi.org/10.30998/sap.v2i2.2028

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory library*. SAGE Publications.

Hammersley, J. M., & Handscomb, D. C. (1964). *Monte Carlo methods*. Springer Dordrecht. https://doi.org/10.1007/978-94-009-5819-7

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

Hikamudin, E. (2017). Estimasi kemampuan siswa dalam ujian nasional menggunakan metode Bayes. *Jurnal Penelitian Kebijakan Pendidikan, 10*(2), 1-14, https://doi.org/10.24832/jpkp.v10i2.171

Insuk, K. (2007). *A comparison of a Bayesian and Maximum Likelihood algorithms for estimation of a multilevel IRT model*. Doctoral dissertation, The University of Georgia, Athens, Georgia. https://openscholar.uga.edu/record/8523/files/kim_insuk_200705_phd.pdf

Lê, S., Josse, J., & Husson, F. (2008). FactoMineR: An R package for multivariate analysis. *Journal of Statistical Software*, *25*(1), 1–18. https://doi.org/10.18637/jss.v025.i01

Mahmud, J. (2017). Item Response Theory: A basic concept. *Educational Research and Reviews*, *12*(5), 258–266. https://doi.org/10.5897/err2017.3147

Mahmud, J., Sutikno, M., & Naga, D. (2016). Variance difference between maximum likelihood estimation method and expected a posteriori estimation method viewed from number of test items. *Educational Research and Reviews*, *11*(16), 1579–1589. https://academicjournals.org/journal/ERR/article-abstract/B2D124860158

Retnawati, H. (2014). T*eori respons butir dan penerapannya: Untuk peneliti, praktisi pengukuran dan pengujian, mahasiswa pascasarjana*. *Nuha Medika.*

Retnawati, H. (2015). Perbandingan estimasi kemampuan laten antara metode maksimum likelihood dan metode Bayes. *Jurnal Penelitian dan Evaluasi Pendidikan*, *19*(2), 145–155. https://doi.org/10.21831/pep.v19i2.5575

Yendra, R., & Noviadi, E. T. (2015). Perbandingan estimasi parameter pada distribusi eksponensial dengan menggunakan Metode Maksimum Likelihood dan Metode Bayesian. *Jurnal Sains Matematika dan Statistika, 1*(2), https://doi.org/10.24014/JSMS.V1I2.1960